DOCUMENT RESUME

ED 312 276 TM 014 017

TITLE	Sex and Race Differences on Standardized Tests. Oversight Hearings before the Subcommittee on Civil and Constitutional Rights of the Committee on the Judiciary. House of Representatives, One Hundredth Congress, First Session.
INSTITUTION	Congress of the U.S., Washington, D.C. House Committee on the Judiciary.
PUB DATE	23 Apr 89
NOTE	309p.; Serial No. 93. Portions contain small/semi-legible print.
AVAILABLE FROM	Superintendent of Documents, Congressional Sales Ofice, U.S. Government Printing Office, Washington, DC 20402.
PUB TYPE	Legal/Legislative/Regulatory Materials (090) -- Reports - Research/Technical (143)
EDRS PRICE	MF01/PC13 Plus Postage.
DESCRIPTORS	Academically Gifted; Admission (School); Classification; *College Entrance Examinations; Elementary Secondary Education; Equal Education; Equal Opportunities (Jobs); Females; Handicap Discrimination; Hearings; Higher Education; High Risk Students; Minority Groups; Occupational Tests; Politics of Education; *Racial Differences; *Sex Differences; *Standardized Tests; *Test Bias
IDENTIFIERS	Admissions Testing Program; Congress 100th; Educational Testing Service; *Scholastic Aptitude Test

ABSTRACT

The purpose of this 1-day hearing was to assess the level and effects of bias based on gender and race differences affecting standardized tests. The focus was on examining the role of standardized tests with respect to educational and employment opportunities for women and minorities. Testimony or statements from 14 witnesses are presented. Subjects addressed include the influence of test scores on entry into gifted programs, uses of the Scholastic Aptitude Test, the impact of standardized testing on children at risk, misclassification of minority students, testing of the handicapped, politics of testing, the Admissions Testing Program of the College Board, test fairness assurances, and the Educational Testing Service's sensitivity review process and its standards for quality and fairness. (TJH)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED312276

TM

SEX AND RACE DIFFERENCES ON STANDARDIZED TESTS

OVERSIGHT HEARINGS

BEFORE THE

SUBCOMMITTEE ON
CIVIL AND CONSTITUTIONAL RIGHTS

OF THE

COMMITTEE ON THE JUDICIARY
HOUSE OF REPRESENTATIVES

ONE HUNDREDTH CONGRESS

FIRST SESSION

ON

SEX AND RACE DIFFERENCES ON STANDARDIZED TESTS

APRIL 23, 1987

Serial No. 93

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it

☐ Minor changes have been made to improve reproduction quality

- Points of view or opinions stated in this document do not necessarily represent official OERI position or policy

Printed for the use of the Committee on the Judiciary

74-668

U.S. GOVERNMENT PRINTING OFFICE
WASHINGTON : 1989

For sale by the Superintendent of Documents, Congressional Sales Office
U.S. Government Printing Office, Washington, DC 20402

BEST COPY AVAILABLE

COMMITTEE ON THE JUDICIARY

PETER W. RODINO, Jr., New Jersey, *Chairman*

JACK BROOKS, Texas
ROBERT W. KASTENMEIER, Wisconsin
DON EDWARDS, California
JOHN CONYERS, Jr., Michigan
ROMANO L. MAZZOLI, Kentucky
WILLIAM J. HUGHES, New Jersey
MIKE SYNAR, Oklahoma
PATRICIA SCHROEDER, Colorado
DAN GLICKMAN, Kansas
BARNEY FRANK, Massachusetts
GEO. W. CROCKETT, Jr., Michigan
CHARLES E. SCHUMER, New York
BRUCE A. MORRISON, Connecticut
EDWARD F. FEIGHAN, Ohio
LAWRENCE J. SMITH, Florida
HOWARD L. BERMAN, California
RICK BOUCHER, Virginia
HARLEY O. STAGGERS, Jr., West Virginia
JOHN BRYANT, Texas
BENJAMIN L. CARDIN, Maryland

HAMILTON FISH, Jr., New York
CARLOS J. MOORHEAD, California
HENRY J. HYDE, Illinois
DAN LUNGREN, California
F. JAMES SENSENBRENNER, Jr., Wisconsin
BILL McCOLLUM, Florida
E. CLAY SHAW, Jr., Florida
GEORGE W. GEKAS, Pennsylvania
MICHAEL DeWINE, Ohio
WILLIAM E. DANNEMEYER, California
PATRICK L. SWINDALL, Georgia
HOWARD COBLE, North Carolina
D. FRENCH SLAUGHTER, Jr., Virginia
LAMAR S. SMITH, Texas

M. Elaine Mielke, *General Counsel*
Arthur P. Endres, Jr., *Staff Director*
Alan F. Coffey, Jr., *Associate Counsel*

Subcommittee on Civil and Constitutional Rights

DON EDWARDS, California, *Chairman*

ROBERT W. KASTENMEIER, Wisconsin
JOHN CONYERS, Jr., Michigan
PATRICIA SCHROEDER, Colorado
CHARLES E. SCHUMER, New York

F. JAMES SENSENBRENNER, Jr., Wisconsin
MICHAEL DeWINE, Ohio
WILLIAM E. DANNEMEYER, California

Catherine A. LeRoy, *Counsel*
Alan Slobodin, *Associate Counsel*

CONTENTS

WITNESSES

Page

Phyllis Rosser, Contributing Editor, Ms. Magazine; Nancy S. Cole, Dean, College of Education, University of Illinois; and Diana Pullin, Associate Dean, College of Education, Michigan State University 2

Statement of Phyllis Rosser 6

Statement of Nancy S. Cole 32

Statement of Diana Pullin 40

Gretchen W. Rigol, Executive Director, Access Services, the College Board; and Carol Anne Dwyer, Executive Director for Test Development, School and Higher Education Programs, Educational Testing Service 69

Statement of Gretchen W. Rigol 74

Statement of Carol Anne Dwyer 151

Michael C. Behnke, Director of Admissions, Massachusetts Institute of Technology; and Denise Carty-Bennia, Professor of Law, Northeastern University, and Executive Chair, Fair Test, Boston, MA 281

Statement of Michael C. Behnke 285

Statement of Denise Carty-Bennia 297

SEX AND RACE DIFFERENCES ON STANDARDIZED TESTS

THURSDAY, APRIL 23, 1987

House of Representatives,
Subcommittee on Civil and Constitutional Rights,
Committee on the Judiciary,
Washington, DC.

The subcommittee met, pursuant to call, at 9:33 a.m., in room 2226, Rayburn House Office Building, Hon. Don Edwards (chairman of the subcommittee) presiding.

Present: Representatives Edwards, Schroeder, and Sensenbrenner.

Staff present: Catherine LeRoy, chief counsel; Alan Slobodin, associate counsel; Barbara Dobynes-Ward, clerical staff.

Mr. **Edwards**. The subcommittee will come to order.

The gentleman from Wisconsin.

Mr. **Sensenbrenner**. Mr. Chairman, I ask unanimous consent that the subcommittee permit coverage of this hearing, in whole or in part, by television broadcast, radio broadcast or still photography, in accordance with Committee Rule 5.

Mr. **Edwards**. Without objection, so ordered.

The purpose of today's hearing is to examine the role of a variety of standardized tests with respect to educational and employment opportunities for women and minorities.

Americans, especially students, are forced to take an increasing number of standardized tests. These tests are used for purposes of school admittance, placement and graduation. Because decisions affecting educational opportunities and employment opportunities are based on these test results, we need to know that educational and vocational tests are, in fact, valid measurements of ability.

The courts in California recently banned the administration of any IQ test to black students when it was found that the tests were biased. When a test scores students on the basis of their race and not their ability, then clearly the test should not be used. Tests that measure culture in the name of ability deny students and workers equal access to employment and educational opportunities. On the basis of public policy and simple fairness, we need to know where test biases exist and what steps can be taken so that they can be eliminated.

Our witnesses will be appearing on three panels. I recognize the gentleman from Wisconsin, Mr. Sensenbrenner.

Mr. **Sensenbrenner**. Mr. Chairman, the minority has no opening statement this morning.

Mr. **Edwards**. Thank you, Mr. Sensenbrenner.

The members of our first panel are Ms. Phyllis Rosser, Contributing Editor, Ms. Magazine; Dr. Diana Pullin, Associate Dean, College of Education, Michigan State University, Lansing; and Dr. Nancy Cole, Dean, College of Education, University of Illinois, Champaign-Urbana, IL. If the members will come to the witness table, please, we will start with the two who are here.

Would you raise your right hands, please. Do you solemnly swear or affirm that the testimony you are about to give is the truth, the whole truth, and nothing but the truth?

Ms. **Rosser**. I do.

Dr. **Cole**. I do.

Mr. **Edwards**. Welcome. Without objection, all the statements will be made part of the record. I believe that Phyllis Rosser is first. Ms. Rosser, as I said, is a contributing editor to Ms. Magazine.

STATEMENTS OF PHYLLIS ROSSER, CONTRIBUTING EDITOR, MS. MAGAZINE; NANCY S. COLE, DEAN, COLLEGE OF EDUCATION, UNIVERSITY OF ILLINOIS; AND DIANA PULLIN, ASSOCIATE DEAN, COLLEGE OF EDUCATION, MICHIGAN STATE UNIVERSITY

Ms. **Rosser**. Thank you. I am glad to be here and pleased that the subcommittee is focusing attention on this important issue.

My name is Phyllis Rosser and I'm a consultant on sex bias in testing. As a contributing editor to *Ms.* magazine for the past 14 years, I have had articles on education and testing published in *Ms.* and other magazines as well. I began researching sex bias in testing in 1979 and wrote a report for *Ms.* at that time on the aptitude tests used for college and graduate school admissions, on standardized achievement tests that are given from kindergarten through the 12th grade for tracking, on IQ tests that are administered by psychologists, and on interest inventories that are used for career guidance in high school.

Most recently, I have been working with the National Center for Fair and Open Testing and am principal author of *Sex Bias in College Admissions Tests: Why Women Lose Out.*

The tests where sex bias seems to have the greatest impact on girls' educational opportunities are the college admissions exams. The Scholastic Aptitude Test and the Preliminary Scholastic Aptitude Test/National Merit Qualifying Test, published by Educational Testing Service and the American College Testing Program's Assessment ACT, are taken by over three million students each year. They are systematically underpredicting the abilities of high school girls. Although females have higher grades than males in all subjects in high school and higher college grades, even the freshman year, senior high school girls averaged 61 points lower than boys on the SAT last year, 50 points lower on the math section, and 11 points lower on the verbal. This is an area where girls excelled until 1972. Then boys began to outscore them and the scope gap has gradually widened.

ETS's justification for the use of this test is that it predicts freshman year college grades, but it is not doing that for girls. They are

52 percent of the 1.5 million test takers, so this means that scores are being underpredicted for approximately 800,000 females every year. In fact, if these tests were accurately predicting freshman grades, girls would score 20 points higher than boys, rather than 61 points lower.

I am sure, if boys were receiving higher grades and lower test scores, the tests would be rewritten.

Minority females are doubly penalized by the test. They all score lower than the males in their ethnic group who, in turn, score lower than white males. In 1985, black women scored 43 points lower than black men, and 264 points lower than white men.

A similar pattern of test bias can be found on ETS's Preliminary Scholastic Aptitude Test/National Merit Qualifying Test, taken by 1.1 million juniors last year, 54 percent of whom were female. Girls' score averages were 53 points lower, in SAT terms, than boys, 41 points in the math, and 12 points in the verbal. To qualify for the National Merit Scholarship, verbal scores are doubled and the math is added, in order to give girls more of a chance. But doubling their lower verbal scores now works against them.

Girls are also scoring lower on the ACT Assessment, taken annually by nearly a million high school seniors, and 54 percent of them are also females. Last year, girls scored lower than boys in math usage, natural science and social studies, but slightly higher in English usage, averaging six score units lower than boys on the test overall.

Sex bias on these tests is having a much greater impact on females than we realize. By underpredicting their academic performance, these tests affect girls chances to gain entrance to nearly 1,500 colleges and universities that require SAT scores or use SAT cut-off scores for admission.

For instance, the University of Texas at Austin requires a combined SAT score of 1,100 for out-of-state applicants. The University of California at Berkeley adds the SAT and three achievement test scores—also tests where the girls score lower—to the student's grade point average, which is multiplied by 1,000 in order to rank candidates for admission.

Unfairly low test scores also become a self-fulfilling prophecy, causing girls to lower their expectations and apply to less competitive schools than their grades suggest. This is truly unfortunate. MIT has been accepting girls with lower SAT math scores and has found they are doing just as well as boys in freshman math classes.

High school girls are also being denied the opportunity to take academic enrichment programs and accelerated courses offered to students with high test scores. A number of summer programs are offered publicly by States or privately by Ivy League and other competitive schools and by well-known prep schools.

Use of these tests also means less scholarship money for female college students. Merit scholarships awarded by hundreds of corporations, foundations, Government agencies, professional organizations and unions each year are partially based on ACT, PSAT, or SAT scores. Most of these organizations refuse to provide a gender or racial breakdown of recipients. However, the National Merit Scholarship Corporation, which offers the most prestigious awards for academic excellence, publishes this data. Over $23 million provided by 670 corporations, foundations, professional organizations, colleges and universities, is given annually by National Merit to students with the highest scores on the PSAT. Last year, girls' qualifying scores averaged 65 points lower than boys, in SAT terms, and they received only 36 percent of the 6,026 available scholarships, while boys received 64 percent. This year, the semifinalists pool, based solely on PSAT scores from which the winners will be chosen, is 34.7 percent female, 61 percent male, and the sex of 4.3 percent is unknown.

In the escalating competition for top students, merit scholarships are being increasingly used for recruitment. Students with high scores on the SAT or the PSAT receive letters offering honor scholarships from a large number of colleges and universities, which buy their scores and other student data from ETS.

The final result of all of this is a real dollar loss for females in later life, as they get less prestigious jobs, earn less money, and have fewer leadership opportunities. Of course, the life-long loss of self-confidence can't be measured in financial terms.

At present, researchers cannot easily tell which questions are biased by examining the tests. Only the test publishers know which questions females and minorities answer incorrectly, and they have not made this information easily available. But there are some theories about the gender gap, particularly on the SAT.

ETS President Gregory Anrig says that a larger pool of test takers will have lower scores. ETS also says that more girls than boys from lower income families take the test. They also have lower test scores which reduces the female average. However, despite their larger pool and lower incomes, the girls who took the SAT in 1985, according to College Board data, had higher grades than the boys who took it. This test didn't reflect their performance in the classroom. ETS says girls take less math and science in high school, but College Board data for 1985 shows that girls who take the test are almost as likely as boys to have taken four years of math.

The College Board says men take harder courses in college, but their own validity studies show girls college grades in math and engineering tend to be underpredicted by their SAT scores.

Most insidious of all are those who say girls' grades reflect good classroom behavior rather than high intelligence. Of course, grades include much ore than can ever be measured on a multiple-choice test, such as the ability to think complexly, solve problems, organize information and express oneself clearly. It is generally acknowledged that girls write better, and the writing tests bear this out.

I have looked at SAT questions over the years and find them offensive in their consistent male orientation. I recently analyzed 24 reading comprehension passages that appeared on four SAT's given in the 1984–85 year. I found references to 42 men and three women in the 24 passages. Thirty-four of these men were famous and their work was cited. One famous woman, Margaret Mead, was mentioned, and her work was criticized.

David White, a lawyer from California, who has done considerable research on college admission exams, has found a number of questions that are demeaning and emotionally loaded for women and minorities. One question on the law school admissions test, published by ETS, concludes that "children should be raised only by their mother and not farmed out to day care centers and full-time babysitters." Certainly women who take this test are going to respond differently to this language than men. It may slow them down and even shake their confidence for a while.

ETS could change these tests to make them fairer, but has chosen not to do so. The Stanford-Binet IQ test is written with the assumption that the sexes are equally intelligent, and it is revised periodically to keep them equal. ETS receives $17,250,000 for the SAT every year, so it could easily afford to change it, to make it sex-fair.

Recent research indicates that other tests are also biased, such as the standardized achievement tests used for high school tracking and the Armed Services Vocational Aptitude Battery, widely used for career guidance in high schools.

I would like the Congress to request that the Department of Education investigate tests that are having major impacts on students, to see if they predict what they are supposed to. In order to do this fairly and accurately, I think it is essential that the researchers who receive these contracts are not connected with the test publishers.

I have additional supporting material that I would like to include with my testimony, and I would like your permission to do that.

Mr. EDWARDS. Without objection, so ordered.

Ms. ROSSER. Thank you.

[The statement of Phyllis Rosser, with attachments, follow:]

TESTIMONY OF PHYLLIS ROSSER

TO THE HOUSE JUDICIARY COMMITTEE ON CIVIL AND CONSTITUTIONAL RIGHTS

April 23, 1987

I have been a Contributing Editor to Ms. Magazine for the past fourteen years and I've had many articles on education and learning published in other magazines as well. I began researching sex bias in testing for Ms. in 1979, with an open mind. Tests had never kept me from anything I wanted to do. I don't even remember the SAT scores I received in 1951.

I examined the tests, read testing studies and interviewed the testing researchers who had written them. I wrote a report for Ms. in 1980 on Aptitude Tests that are used for college and graduate school admissions, Standardized Achievement Tests given from kindergarten through 12th grade, I.Q. tests administered by psychologists, and Interest Inventories used for Career Guidance in high school.

I am very pleased that Congress is interested in the effects standardized tests are having on females and sorry to report that the tests have not improved much since I began my research. In fact, on the college entrance examinations, the score gap between the sexes has widened.

What struck me first when I looked at these tests was the overwhelming number of males that populated them - all of whom were engaged in traditional occupations like doctor and lawyer while women were teachers, nurses and secretaries. According to recent research, there are still twice as many men as women on most tests and they are still shown in stereotyped roles, even though this doesn't represent the world of 1987 at all. Studies done by Educational Testing Service researchers as far back as 1979 ("Sex Differences and Sex Bias in Test Content" by Ekstrom, Lockheed, Donlon, Educational Horizons) show that "females tend to do better on items that have more female or neutral figures than on items in which there are male figures." This means that male-oriented content is not only offensive, it is also a source of bias.

But the tests where sex bias seems to have the greatest impact on girls' educational opportunities are the college entrance examinations. The Scholastic Aptitude Test (SAT) and the Preliminary Scholastic Aptitude Test/National Merit Qualifying Test (PSAT/NMSQT) published by Educational Testing Service and the American College Testing Program's ACT Assessment (ACT) are systematically underpredicting the abilities of high school girls. Although females have higher grades in every subject in high school and higher college grades, they receive lower test scores on the SAT and the ACT.

The SAT is composed of two sections, Verbal and Math, each scored on a 200-800 point scale. The maximum possible score is 1600. Last year, women's average SAT scores were 61 points lower than men's - 50 points on the Math section and 11 points on the Verbal section - an area where girls excelled until 1972. Then boys began to outscore them verbally as well as mathematically (boys have always received higher math scores on this test), and the score gap has gradually widened.

This growing score gap is surprising since ETS says the main purpose of this test is to predict freshman year grades but it's not doing that for girls. They make-up 52% of the 1.5 million test takers so this means that scores are being underpredicted for approximately 800,000 females every year. If this test were accurately predicting freshman year grades, girls would score 20 points higher than boys rather than 61 points lower.

I'm sure, if boys were receiving higher grades and lower test scores, the tests would be rewritten.

Minority women are doubly-penalized by the test. They all score lower than the men in their ethnic group, who, in turn, score lower than white men. In 1985, black women scored 43 points lower than black men and 264 points lower than white men.

A similar pattern of test bias can be found on ETS' Preliminary Scholastic Aptitude Test/ National Merit Qualifying Test (PSAT/NMQT), taken by 1.1 million junior high school students last year (who were 54% female). ETS promotes this as a practice test for the SAT, but the National Merit Scholarship Corporation awards over $23 million in student scholarships to the highest scorers on this test.

Like the SAT, the PSAT/NMSQT has two parts. Each is scored on a scale of 20-80. Testmakers claim an approximation of future SAT scores can be obtained by multiplying PSAT/NMSQT scores by ten. In 1985-86, girl's score averages were 53 points lower, in SAT terms, than boys: 41 points in the Math, 12 points in the Verbal. To qualify for the National Merit Scholarship, verbal scores are doubled and the math is added - in order to give girls more of a chance. But their lower verbal scores which are doubled, are now working against them.

An alternative college entrance exam to the SAT is the ACT Assessment, a survey achievement test taken annually by nearly a million high school seniors (54% of whom are female), mainly in the Mid-West, Southwest and South. The ACT has four sections: English Usage, Mathematics Usage, Natural Science, And Social Studies. The test is scored on a scale that ranges from 1-36. In 1985-86, girls averaged 2.8 score units lower than boys in Math Usage, 2.5 units lower in Natural Science, and 1.7 units lower in Social Studies but slightly higher (1.0 units) in English Usage, averaging 6 units lower than boys on the test, overall.

Girls also receive lower scores on most of the Achievement Tests published by ETS which are required for admission to some colleges and universities. According to the College Board's *Profiles of College-Bound Seniors*, 1985, girls scored nine points higher on English Composition and Literature, one point higher on German but lower on all the other tests.

Sex bias on these tests is having a much greater impact on females than we realize. By underpredicting their academic performance, these tests affect girls chances to gain entrance to colleges and universities that require SAT or ACT scores, or use them as cut-off scores for admission. They also markedly diminish their chances to obtain merit scholarships based on test scores, and to enter many special educational programs for gifted high school students that use SAT scores in their admissions criteria.

Test Scores Effect College Admission

Nearly all the 1500 accredited colleges and universities in the country require students to submit SAT or ACT scores for admission. Some use them as cut-off scores and others put them into an admissions formula. (see appendix I for a list of the colleges and uiversities requiring cut-off scores or using SAT scores as part of a numerical formula.) For example, the University of Texas at Austin requires out-of-state applicants to have minimum SAT scores of 1100. The University of California at Berkeley adds the SAT score, the scores on three ETS Achievement Tests (where girls also receive lower scores) and the Grade Point Average multiplied by 1000, to rank candidates for admission.

Although some colleges may not actually use scores in the selection process, they often publish the average SAT scores of their previous freshmen class to establish high academic credentials. As a result, women with lower SAT scores will lower their expectations and apply to less competitive schools than their grades suggest. Ernest Boyer recently reported in College: The Undergraduate Experience in America that 62% of the students questioned said they lowered their college expectations after receiving their SAT scores.

Low Test Scores Reduce Entry Into 'Gifted' Programs

A large number of academic enrichment programs are offered to students with high SAT or PSAT scores. Fewer of these opportunities are offered to females, due to their lower scores. This means they not only lose the opportunity to enhance or accelerate their high school program, but also have less impressive resumes of extracurricular academic activities to present on college applications.

In New Jersey, outstanding honors students in science and political science with high SAT scores are invited to attend the Governor's School, a summer enrichment program held on college campuses. 65% of the attendees at the science school this summer will be male, 35% will be female, from a pool of applicants that was 75% male. High PSAT scores and high grade point averages are also used to select one student from each high school in New Jersey to attend the New Jersey Scholars Program held at the Lawrenceville School each summer.

In Washington, D.C., students with high SAT Math scores are offered opportunities to take advanced math courses on college campuses during the summer. Additionally, high scoring students whose parents can afford summer school tuition have a smorgasbord of opportunities to develop their giftedness. Summer enrichment courses are offered by Ivy League and other competitive schools, and by well-known prep schools. This summer, the George School in Newton, Pennsylvania and Blair Academy in Blairstown, New Jersey will offer courses in advanced mathematics, college science, computer science, languages, literature, the arts and, ironically, PSAT and SAT coaching.

Johns Hopkins University's Center for the Advancement of Academically Talented Youth (CTY) invited 26,876 seventh grade boys and girls in 19 states to take the SAT, to determine if they were mathematically or verbally talented. Junior High School students qualify for this by scoring in the upper 3% on the mathematics section of a national standardized achievement test. Those who score 500 or more on the Verbal or Math section are invited to attend one of their five camps for "gifted and talented" students.

This summer, invitations to the Johns Hopkins program will be extended to over 2,500 boys but only 1,081 girls. Although an equal number of boys and girls take the test, girl's lower SAT scores keep them from qualifying for these high-powered summer programs. They may also suffer a blow to their self esteem and lower their expectations about future SAT performance - before they even reach high school.

Low Test Scores Deny Merit Scholarship Money

Use of exam scores also means less merit scholarship money for female college students. Merit scholarships awarded by hundreds of corporations, foundations, government agencies, professional organizations and unions each year are partially based on ACT, SAT or PSAT scores. Most of these organizations refuse to provide a gender or racial breakdown of scholarship recipients. However, the National Merit scholarship Corporation, which offers the most prestigious awards for academic excellence, publishes this data.

Over 23 million dollars, provided by 670 corporations, foundations, colleges and universities are given annually to students with the highest PSAT scores. Last year girls qualifying scores averaged 65 points lower than boys (in SAT terms) and they received only 36% of the 6,026 scholarships awarded while boys received 64%. This year the semi-finalist pool (based solely on PSAT scores) from which the winners will be chosen has 15,507 students. 34.7% are female and 61% are male (the sex of 4.3% is unknown). (see State-By-State breakdown, p.10)

Semi-finalist status is given to students whose PSAT scores (twice Verbal and Math score) rank them in the top half of 1% in each state. In order to obtain scholarship money, semi-finalists submit information about their academic records, extracurricular activities, leadership potential and intended college major, along with their principal's recommendation to the National Merit Corporation's selection committee. Students must also duplicate their high PSAT score with "an equivalent high Scholastic Aptitude Test performance," according to their Program Guide. This also works against lower-scoring females. In 1985-86, of the 13,777 Merit Finalists, 64.1% were male and 35.9% were female. 43.7% of the finalists actually receive Merit Scholarships.

An alarming trend for women is evident in the National Merit Corporation's Annual Reports. Although the total number of scholarships awarded annually has increased, the number and percentage of female recipients has decreased noticably in the last three years. In 1983-84, National Merit Scholars were 40.2% female; in 1984-85, 37.9% were female; in 85-86, 36% were female.

It is impossible to calculate exactly how many millions of dollars girls lost in this uneven split because Merit Scholarships are awarded in three catagories. National Merit Corporation awards 1,800 of it's own $2,000 scholarships annually. In addition, it administers the awarding of scholarships for 425 corporations and 2,800 colleges and universities in amounts ranging from $250 to $8,000 per year.

The National Merit Corporation also administers the awarding of 1,179 "Special" corporate scholarships worth $7.6 million. These scholarships are awarded to students with scores below the finalist level who are interested in a career the grantor wants to encourage, or who live in a community where the company has offices. (see Appendix II for list of corporate and business sponsors of special merit scholarships in 1986)

New York State's Merit Scholarships, worth over $40 million annually, are awarded to students who have the highest ACT or SAT scores in each of New York's counties. In 1986-87, 672 of the 1,000 Empire State Scholarships of Excellence awarded were to boys, while only 270 went to girls. The gender of

58 winners could not be determined by name.

Males also won more of New York State's 25,000 Regents College Scholarships, which are exclusively determined by SAT or ACT scores, and worth up to $1,250 each. Of the 109,266 students who competed for the scholarships, 47% were male and 53% were female. However, 57% of the 25,277 winners were male and 43% were female.

Since FairTest and NYPIRG made this discrepancy public, the New York State Board of Regents moved swiftly. Acknowledging that women's lower SAT scores kept them from receiving their fair share of merit scholarships, the Regents voted unanimouly to ask the legislature for funds to develop a new, unbiased tests.

Other states use a combination of grades and test scores for their merit programs with more equitable results. New Jersey requires students to have SAT scores of 1200 or more and also rank in the top 10% of their high school class to qualify for Garden State Distinquished Scholarships. Up to $4,000 is awarded annually to 800 students (at least 2 from each school) for a total of $3,200,000 to encourage them to attend colleges in New Jersey. Last year's Garden State Distinguished Scholars were 50% female and 50% were male.

A computer print-out from a typical northeastern high school guidance office lists 134 scholarships tied to test scores. These "merit" scholarships are given by unions, fraternal organizations, religious denominations, corporations (mainly sponsoring children of employees), professional organizations, and the military. Most of these scholarships are awarded to students with high test scores in combination with high grades, an interest in pursuing a particular couse of study and/or financial need. Engineering societies predominate, giving more career-based merit scholarships than any other group. (see Appendix III for a partial listing of private scholarships based on SAT and ACT scores)

In the escalating competition for top students, merit scholarships are being increasingly used for recruitment, according to a 1984 study, more than 85% of four-year private colleges and nearly 90% of public institutions offer no-need scholarships for academic excellence, and substantially more of these are being offered now than even five years ago. In private, four-year colleges, 44% of this no-need money is taken from tuition and fee income, raising important questions about the spiraling costs of college tuition.

Last year, one New Jersey student who received a $4,000 Garden State Distinquished Scholarship, found his mailbox full of additional scholarship offers. Thirteen New Jersey colleges offered him grants ranging from $2,000 to $12,000. Drew University in Madison, N.J. also told him that it offers $48,000 to students who score 1350 or better on the SAT and $32,000 to students with 1300 SAT's.

Two out-of-state colleges offered this student "honors" scholarships outright, ranging from $500 to $10,000. Sixteen other colleges and universities told him he qualified for their merit scholarships, some of which covered full tuition. In addition, eight universities - including the Universities of Michigan, Indiana, and Delaware - offered him admission to their Honors Programs in which a small, select group of academically-talented students attend a smaller, select college within the university. They are given enriched academic programs, honors grants, and live together in a separate residence hall.

The final result of lower test scores is a real dollar loss for females in later life as they get less prestigious jobs, earn less money, and have fewer leadership opportunities. Of course, the life-long loss of self-confidence can't be measured in financial terms.

Why The Gender Gap?

It is impossible to tell which questions are biased by examining the tests. Only the test publishers know which questions females and minorities answer incorrectly and they have not made this information easily available. A bill is currently moving through the New York State Legislature which would require publishers to provide a gender and racial analysis of test questions for an entire year.

In the meantime, there are some theories about the gender gap, particularly on the SAT.

ETS President Gregory Anrig says that a larger pool of test takers will have lower scores. ETS also says that the larger pool of girls includes more girls from lower income families who have lower test scores which in turn reduces the average female scores. However, the girls who took the SAT in 1985, according to the College Board's Profiles of College-Bound Seniors, had higher grades than the boys who took it, despite their larger pool and lower incomes.

Fred Marino, Assistant Director of Public Affairs for The College Board, says "girls take less math and science in high school than boys," to explain the 50 point gap on the math section. However, the College Board's Profiles for 1985 shows that girls who take the test are almost as likely as boys (50.5% vs. 57.6%) to have taken four years of math.

He also says that girls take easier courses in college. They are less likely to be taking science and engineering where grades are lower because the courses are harder. But The College Board's own validity studies show that women who major in engineering and math in college tend to receive higher grades than their SAT scores had predicted. Massachusetts Institute of Technology has been admitting women with lower SAT Math scores and finds they do just as well as men in freshman math classes.

ETS also says the tests reflect the bias against females in society. They suggest that girls are treated differently in the classroom which may effect the way they perform on standardized tests. Although the society is biased against females and the classroom reflects that, girls are able to overcome this handicap and earn better grades, even though they receive less classroom attention.

Most insidious of all are those who say girls' grades reflect good classroom behavior rather than high intelligence. As we all know, grades include much more than can be measured on a multiple-choice test, such as the ability to think complexly, solve problems, organize information and express oneself clearly. It is generally acknowleged that girls write better, and the writing tests even bear this out.

I have looked at SAT questions over the years and find them offensive in their consistent male orientation. I recently analyzed 24 reading comprehension passages that appeared on 4 SAT's given in the 1984-85 year. (There are six reading passages on each SAT.) I found references to 42 men and three women in the 24 passages. 34 of these men were famous and their work was cited. One women, Margaret Mead, was famous and her work was criticized.

David White, a lawyer from California, has done considerable research on the graduate entrance exams published by ETS - the Law School Admissions Test (LSAT), the Graduate Record Examination (GRE) and The Graduate Management - Admissions Test (GMAT) - and found a number of questions that are emotionally-loaded and offensive for women and blacks. For example, one question on the LSAT concludes that "children (should) be raised only by their mothers, and...not be farmed out to day-care centers and full-time babysitters." Certainly the mothers who are taking this test are going to be "farming out their children."

David White and I both feel this type of demeaning question slows down test takers *and* may even shake their confidence for a while on a test that requires the utmost in speed and risk-taking. At the very least, they cast doubt on ETS's Sensitivity Review.

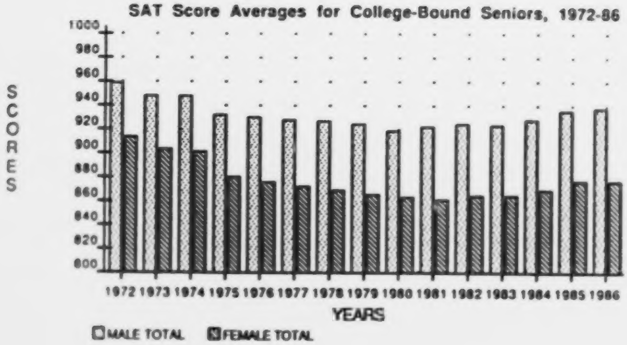
ETS could change these tests to make them fairer for girls but has chosen not to do this. The widely-used Stanford-binet I.Q. Test is written with the assumption that the sexes are equally intelligent. It is periodically revised to make sure the sexes score equally well.

I would like to ask ETS why it has decided that boys are smarter than girls? I would also like to know what the SAT is predicting, if it's not freshmen grades? ETS receives $17,250,000 for this test every year that doesn't do what it's supposed to for over half the people taking it. I think that is consumer fraud.

I also think unfair college admissions tests may be the tip of the iceberg. Recent research indicates that other tests are also biased against girls, like the standardized achievement tests used for high school tracking and the Armed Services Vocational Aptitude Battery (ASVAB), the most widely-used aptitude test for career guidance in high schools.

I would like Congress to request that the Department of Education investigate tests that are having major impacts on students - to see if they predict what they are supposed to. In order to do this fairly and accurately, I think it is essential that the researchers who receive these contracts are not connected with the test publishers.

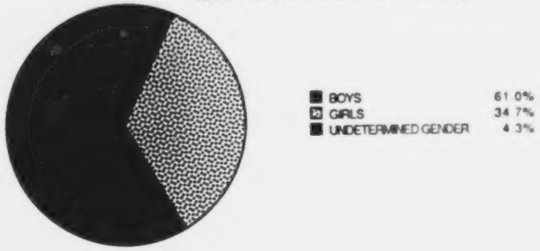
The statistics, charts and some of the information presented here were first published in the National Center for Fair and Open Testing Report on *Sex Bias in College Admissions Tests: Why Women Lose Out* by Phyllis Rosser with the Staff of National Center for Fair and Open Testing, April 1987.


SAT Score Averages for College-Bound Seniors, 1972-86

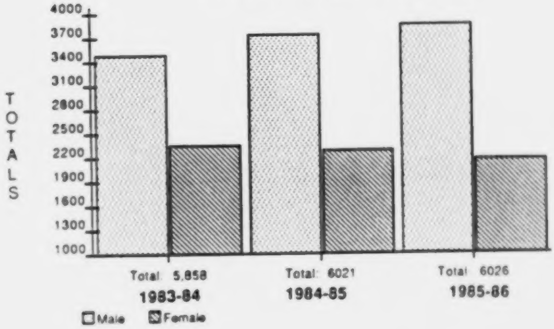
Average SAT Scores	Females	Males	Diff.
Asian-Pacific Americans	897	946	-49
Black	705	748	-43
Mexican-American	775	845	-70
Native Americans	790	855	-65
Puerto Rican	744	820	-76
White	912	969	-57
National Average	877	938	-61

--from 1985 Profiles, College Bound Seniors,
by Leonard Ramist and Solomon Arbeiter, CEEB 1986.

1986 National Merit Scholarship Semi-Finalists



National Merit Scholarship Winners
(over last three years by gender)



National Merit Semifinalists for 1986-87

State by State Breakdown by Gender

STATE:	# GIRLS:	% GIRLS:	# BOYS:	%BOYS:	# UNKNOWN:	TOTAL:
Alabama	82	36.4%	140	62.2%	3	225
Alaska	12	44%	13	48%	2	27
Arizona	55	33.7%	99	60.7%	9	163
Arkansas	47	29.3%	110	68.8%	3	160
California	495	35.4%	817	58.3%	88	1,400
Colorado	70	35%	120	60.6%	8	198
Connecticut	81	33.3%	152	62.5%	10	243
Deleware	10	23.2%	30	69.8%	3	43
D.C.	25	38.5%	39	60%	1	65
Florida	176	32.9%	326	61%	32	534
Georgia	123	36.1%	217	64%	1	341
Hawaii	35	46.7%	34	45%	6	75
Idaho	13	20.6%	49	77.8%	1	63
Illinois	258	33.8%	472	61.9%	33	763
Indiana	134	34.4%	247	63.3%	9	390
Iowa	85	35.6%	148	62%	6	239
Kansas	64	39.7%	86	53.4%	11	161
Kentucky	75	31.9%	153	65.1%	7	235
Louisiana	87	32.4%	168	62.7%	13	268
Maine	28	32.1%	56	64.4%	3	87
Maryland	116	33.9%	206	60.2%	20	342
Massachusetts	178	35.3%	308	61.2%	17	503
Michigan	218	32.4%	426	63.4%	28	672
Minnesota	125	37.1%	203	60.2%	9	337

National Merit Semi-Finalists 1986 (continued)

State:	# GIRLS:	% GIRLS:	# BOYS:	%BOYS:	# UNKOWN:	TOTAL:
Mississippi	44	29.5%	102	68.4%	3	149
Missouri	131	37.8%	192	55.3%	24	347
Montana	32	51.6%	26	41.9%	4	62
Nebraska	39	30.2%	80	62%	10	129
Nevada	19	38%	30	60%	1	50
New Hampshire	45	35.4%	79	62.2%	3	127
New Jersey	170	33%	327	63.5%	18	515
New Mexico	24	26.6%	65	72.2%	1	90
New York	373	32%	730	62.7%	62	1,165
North Carolina	155	39.2%	232	58.7%	8	395
North Dakota	20	41.6%	25	52%	3	48
Ohio	314	39.9%	446	56.7%	27	787
Oklahoma	55	26.6%	138	66.6%	14	207
Oregon	58	36.7%	93	58.9%	7	158
Pennsylvania	303	34.1%	545	61.5%	38	886
Rhode Island	22	33.3%	44	64.7%	2	68
South Carolina	80	34.7%	142	61.7%	8	230
South Dakota	19	39.5%	26	54.2%	3	48
Tennessee	115	38.2%	176	58.5%	10	301
Texas	295	31.8%	579	62.4%	54	928
Utah	41	36.6%	66	58.9%	5	112
Vermont	11	30.5%	24	66.6%	1	36
Virginia	131	35.2%	224	60.2%	17	372
Washington	86	35.5%	135	55.8%	21	242
West Va.	54	40%	79	58.5%	2	135
Wisconsin	111	31.6%	227	64.7%	13	351
Wyoming	13	37.1%	19	54.3%	3	35
TOTAL	5,352	34.7%	9470	61%	685(4.3%)	15507

JUNE 1984 SAT

Reading Comprehension Passage on Margaret Mead

Many social anthropologists and other scientific observers of human communities have emphasized the similarities in the sex roles in various communities. One very distinguished anthropologist, Margaret Mead, in her book *Male and Female*, gives this summary description of the sex roles: "The home shared by a man or men and female partners, into which men bring the food and women prepare it, is the basic common picture the world over. But this picture can be modified, and the modifications provide proof that the pattern itself is not something deeply biological."

It is surprising that Margaret Mead, with her extensive and intensive personal experience of diverse communities throughout the world, should venture upon such a dubious generalization. She is right in describing the preparation of food as a monopoly for women in nearly all communities, but the surmise that the provision of food is a man's prerogative is unwarranted. In fact, an important distinction can be made between two kinds or patterns of subsistence agriculture: one in which food production is taken care of by women, with little help from men, and one in which food is produced by the men with relatively little help from women. As a convenient terminology I propose to denote these two systems as the female and male systems of farming.

33. Which of the following best explains what the author means by "the similarities in the sex roles" (lines 2-3)?

    (A) The equality of men's and women's traditional tasks
    (B) The likenesses in patterns of division of labor between men and women
    (C) The universal acceptance of the need for cooperation between men and women within a community
    (D) The overlapping of tasks performed by men and women in various communities
    (E) The correspondence between a community's attitude toward women and the traditional tasks they perform

34. The author's attitude toward the statement by Margaret Mead is one of

    (A) reluctant consent
    (B) intrigued curiosity
    (C) respectful disagreement
    (D) apologetic defensiveness
    (E) mild endorsement

35. Which of the following best describes the relation between the two paragraphs in the passage?

    (A) The second disputes aspects of the opinions presented in the first.
    (B) The second explains the logic behind the arguments summarized in the first.
    (C) The second provides specific examples of the general statements presented in the first.
    (D) The second questions the social importance of the issues raised in the first.
    (E) The second analyzes the implications for the future of the theories described in the first.

GO ON TO THE NEXT PAGE

Appendix I

Nearly 300 four-year accredited colleges and universities have either absolute cut-off scores or specify a cut-off score as a leading component of their admissions program for all or one of their programs, or utilize test scores in a qualifying numerical formula. The following is the list of those colleges and universities:

Institution:

Abilene Christian University
Akron University
Alabama State University
University of Alabama, Birmingham
Albany State College
Alcorn State University
Allentown College of St. Francis de Sales
Alvernia College
Alma College
Angelo State University
Arizona State University
University of Arizona
Arkansas College
Arkansas State University
University of Arkansas, Fayetteville
Armstrong State College
Auburn University
Augusta College
Austin Peace State University
Avila College
Ball State University
Belmont College
Bemidgi State University
Benedictine College
Bethany College
Bethel College
Black Hills State College
Bluefield State College
Bluffton College
Butler University
California Baptist College
California Polytechnic State University
California State College, Bakersfield
California State College, San Bernadino
California State Polytechnic
California State University, Cinco
California State University, Carson
California State University, Fresno
California State University, Fullerton
California State University, Haywood
California State University, Long Beach
California State University, Los Angeles
California State University, Northridge
California State University, Sacramento
California State University, Turlock
California University of Berkeley
California University of Davis
California University of Irvine
California University of Los Angeles
California University of Riverside
California University of Santa Barbara
California University of Santa Cruz

Cameron University
Castleton State College
Centenary College of Louisiana
Central College
Central Missouri State
Central Florida University
Central State University
Chicago State University
CUNY, Bernard M Barach College
CUNY, Brooklyn
CUNY, City College
CUNY, College of Staten Island
CUNY, Hunter College
CUNY, Queens
Colorado University of Colorado Springs
Colorado University of Denver
Columbia Union College
Concord College
Concordia College
Dakota Wesleyan University
Dana College
Devry Institute of Technology, City of Industry
Devry Institute of Technology, Irving, Texas
Devry Institute of Technology, Decatur, Georgia
Devry Institute of Technology, Chicago
Devry Institute of Technology, Lombard, Illinois
Devry Institute of Technology, Columbus, Ohio
Dickenson State College
East Central University
Eastern Illinois University
East Tennessee State University
East Texas State University
Eastern College
Eastern Kentucky University
Eastern Mennonite College
Eastern New Mexico University
Embry Riddle Aeronautical University
Evansville University
Fairmont State College
Fehcian College
Florida Atlantic University
Florida Southern College
Florida State University
Florida University of Gainesville
Fort Wayne Bible College
Georgia University of Athens
Georgian Court College
Glenville State College
Graceland College
Houston Baptist University
University of Houston, Houston
Humbolt State University
Illinois State University
Indiana State University, Terre Haute
Indiana University, Bloomington
Indiana University, Kokomo
Indiana University, Northwest,Gary
Indiana University, Purdue Univ. at Indianapolis
Indiana University, South Bend
Iowa State University,Ames
Iowa University of Iowa City

Jackson State University
Jacksonville State University
Jacksonville University
John Brown University
Kent State University
Kentucky State Unisversity
Kentucky Weseyan College
La Roche College
Lamar University
Lander College
Lewis-Clark State College
Loras College
Louisiana College
Loyola University
Maine, University of, Fort Kent
Maine, University of, Presque Isle
Mankato State University
Marsfield University of Pennsylvania
Mary Hardin-Baylor, University of Belton, Texas
Maryland, University of, College Park
Maryland, University of, Eastern Shore
Massachusetts, University of, Amherst
Mercyhurst College
McMurry College
Mesa College
Memphis State University
Metropolitan State College, Denver
Miami Christian College
Middle Tennessee State University
Minot State College
Montana College of Mineral Science & Technology
Mississippi College
Mississippi State University
Mississippi, University of
Mississippi, University for Women
Mississippi, Valley State University
Missouri Southern State College
Missouri, University of, Kansas City
Missouri Western State College
Mobile College
Moorhead State University
Molloy College
Morehouse College
Morgan State University
Mcholls State University
Mt. St. Clare College
Mt. Vernon Nazarene College
Nebraska, University of, Lincoln
Nebraska, University of, Omaha
New England, University of, Biddeford
New Mexico Institute of Mining and Technology
New Mexico State University
New York Institute of Technology
New York University
Nichols College
North Alabama, University of
North Arizona University
North Carolina Agricultural and Technical
North Carolina, University of, Asheville
North Carolina, University of, Charlotte
North Central College
Northeast Missouri State University

North Florida, University of, Jacksonville
North Texas State University, Denton
North Arizona University
Northeastern Oklahoma University
Northern Colorado, University of, Greenley
Northern Illinois University
Northern Kentucky University
Northern State College
Northwestern Oklahoma
Northwestern College
Ohio State University
Oklahoma State University
Oklahoma, University of, Norman
Oklahoma Baptist College
Oklahoma Panhandle State University
Old Dommion University
Oregon State University
Oregon, University of, Eugene
Quincy College
Pikeville College
Portland State University
Portland, University of, Portland
Rio Grande College
St. Ambrose College
St. Francis College
St. Cloud University
St. Louis College of Pharmacy
St. Leo College
St. Mary's College of Maryland
St. Mary, College of Omaha, Nebraska
St. Mary's University
St. Paul Bible College
San Francisco State University
San Jose State University
Savannah State College
Science and Arts of Oklahoma, University of Chichaska
Shepard College
Sioux Falls College
Sonoma State University
South Dakota School of Mining and Technology
Southeast Missouri State University
Southern Illinois at Carbondale
Southern Illinois University
Southwest State University
South Dakota State University
Southern Oklahoma University
Southern Colorado, University of
South Nazarene, University of
Southern Oregon State
South West Texas State University
Southwestern Oklahoma State University
South Alabama, University of, Mobile
South Carolina, University of, Aiken
South Carolina, University of, Conway
South Carolina, University of, Spartanburg
South Florida, University of, Tampa
Southeastern Louisiana University
Southern Arkansas University
Southern College of Seventh Day Adventist
Southern Mississippi, University of Hattiesburg
Southern Louisiana, University of Lafayette

Spalding University
Spring Arbor College
Sue Ross University
Stockton State College
SUNY College of Environmental Science and Forestry
SUNY College of Geneseo
SUNY College of New Paltz
SUNY College of Old Westbury
Talladega College
Tarleton State University
Tennesse State University
Tennesse Technological Universirty
Tennesse, University of, Chattanooga
Tennesse, University of, Knoxville
Tennesse, University of, Martin
Texas A and I University
Texas A & M University
Texas College
Texas Tech. University
Texas, University of, Arlington
Texas, University of, Austin
Texas, University of, El Paso
Texas, University of, San Antonio
Thomas Moore College
Toledo, University of
Transylvania University
Trevecca Nazarene College
Trinity College
Tusculum College
Tushesee University
Union University
Valley City State College
United States Merchant Marine Academy
Valley City State College
Virginia, University of, Wise, W. Virginia
Warren Wilson College
Wayne State University
Weber State College
West Liberty State College
West Virginia Institute of Technology
Western Connecticut State University
Western Illinois University
Western Kentucky University
Western Michigan University
Western Oregon State College
Wheeling College
Winona State University
Winston-Salem State University
Wisconsin, University of, LaCrosse
Wisconsin, University of, Kenosha, Parkside
Wisconsin, University of, Superior
Wisconsin, University of, Whitewater
York College of Pennsylvania

Appendix II

List of Corporations who give Special Merit Scholarships, administered by the National Merit Scholarship Program:

Abex Foundation Inc.
Acushnet Foundation
Albany International
Alco Standard Foundation
Allied Van Lines Memorial Fund
Amcast Industrial Foundation
American District Telegraph Company
American Express Foundation
American Optical Foundation
The American Tobacco Company
AmeriTrust Corporation
Ametek Foundation, Inc.
Amfac Inc.
Arthur Andersen & Co. Foundation
Armstrong Rubber Company Fdn., Inc.
Armstrong World Industries, Inc.
The Aro Corporation
Avery International
Bank of America-Giannini Foundation
BASF Corporation Chemicals Division
BASF Corporation-Fibers Division
BASF Corporation Inmont Division
Basic American Foods company
Bell & Howell Foundation
Bemis Company Foundation
Loren M. Berry Foundation
The Black & Decker Corporation
Blue Bell Foundation
The Bristol-Meyers Fund, Inc.
Brockway Glass Company Foundation
Brown & Williamson Tobacco Corporation
Browning-Ferris Industries, Inc.
Burndy Corporation
Burroughs Wellcome Co.
Carson Pirie Scott Foundation
Carter-Wallace, Inc.
Castle & Cooke, Inc.
Celanese Corporation
Central Soya Foundation, Inc.
Centronics Data Computer Corp.
Charter Medical Corporation
Chesebrough-Pond's Inc.
The Clorox Company Foundation
Collins & Aikman Corporation
Combined International Corporation
Combustion Engineering, Inc.
Communications Satellite Corporation
Consolidated Papers Foundation, Inc.
Consolidation Coal Company
The Continental Corporation Foundation
Continental Grain Foundation
Crompton & Knowles Foundation, Inc.
Crum and Forster Foundation
Dart & Kraft Foundation

Data General Corporation
Del Monte Corporation
Diamond Shamrock Corporation
A. B. Dick Company
Dillingham Corporation
R. R. Donnelley & Sons Company
Dow Jones Foundation
Dresser Foundation, Inc.
EDT Group of Penn Central Corporation
The El Paso Natural Gas Company
Equimark Corporation
Estee Lauder Inc.
Ex-Cell-O Corporation
Fafnir Bearing Division of
  The Torrington Company
The Filene Charitable Foundation
Firestone Trust Fund
First Fidelity Bancorporation
First Interstate Bank of Arizona, N.A.
  Educational Foundation
Fischbach Corporation
Fischer & Porter Co.
Florida Steel Corporation Fdn., Inc.
Gannett Foundation, Inc.
General Foods Corporation
Gleason Memorial Fund, Inc.
W.W. Grainger, Inc.
GrandMet USA, Inc.
Great American Insurance Company
Great Northern Nekoosa Fdn., Inc.
Gulf + Western Foundation
Harsco Corporation Fund
Helene Curtis Industries, Inc.
Henkel Corporation
HMI Holdings, Inc.
Hobart Corporation
Hoffmann-LaRoche Inc.
Homestake Mining Company
Geo. A. Hormel & Company
Hospital Corporation of America
Illinois Tool Works Foundation
Insilco Corporation
Interlake Foundation
Ivey Trust Fund
Johnson & Higgins
Johnson Worldwide Associates, Inc.
The Johnson's Wax Fund, Inc.
Kaman Corporation
The Kennametal Foundation
Kenosha Foundation
Kerr-McGee Corporation
Kidde Consumer Durables Corp.
Kidder, Peabody & Co., Incorporated
Knight-Ridder Newspapers, Inc.
Kraft, Inc.
Leeds & Northrup Foundation
Lennox Foundation
Libby, McNeill & Libby, Inc.
The Liberty Corporation Foundation
Thomas J. Lipton Foundation, Inc.
Lloyds Bank California
Loews Foundation

The LTV Foundation
Mattel Foundation
The McGraw-Hill Foundation, Inc.
Estate of John E. McKeen
McKesson Foundation, Inc.
McNeilab, Inc.
Mellon Bank N.A.
Edwin T. Meredith Foundation
Midland-Ross Foundation
Minnesota Mining and Manufacturing Co.
Mitchell Energy & Development Corp.
The Modine Foundation, Inc.
Monsanto Fund
Morgan Guaranty Trust Company
G. C. Murphy Company Foundation
Murphy Oil Corporation
Nabisco Foundation
Nalco Chemical Company
National Distillers Distributors Fdn.
National Medical Enterprises, Inc.
National Starch & Chemical Fdn., Inc.
Nestle Foods Corporation
New Jersey Manufacturers Insurance Co.
Norfolk Southern Foundation
Ortho Pharmaceutical Corporation
Owens-Corning Fiberglas Corporation
Frank E. and Seba B. Payne Foundation
Pechiney Corporation
The Penn Mutual Charitable Trust
PepsiCo Foundation, Inc.
Pet Incorporated
Petrolane Incorporated
Pfizer Inc.
Phelps Dodge Foundation
The Jesse Philips Foundation
PPG Industries Foundation
The Proctor & Gamble Fund
Prudential-Bache Foundation
Public Service Co. of New Hampshire
Puritan-Bennett Corporation
The Quaker Oats Foundation
Quanex Foundation
Quasar/Matsushita Industrial Company
Raytheon & Local 1505 IBEW
RB&W Corporation
The Richman Brothers Foundation, Inc.
The Riegel Textile Corporation Fdn.
RJR Nabisco, Inc.
RKO General, Inc.
St. Joe Minerals Corporation
Sandoz Corporation
Sara Lee Foundation
Schering-Plough Foundation, Inc.
Schlegel Corporation
Service America Corporation
SFN Companies, Inc.
Shaklee Corporation
Shell Companies Foundation, Inc.
Siemens Capital Corporation
Simmonds Precision Products, Inc.
Snap-on Tools Corporation

Robert S. Solinsky Scholarship Fdn.
Sony Corporation of America Fdn. Inc.
The Standard Oil Company
The Stanley Works
State Farm Companies Foundation
Stewart-Warner Foundation
Stranahan Foundation
The Aaron & Lillie Straus Fdn. Inc.
Suburban Propane Gas Corporation
Sun Company, Inc.
Sunshine Biscuits Foundation
Talley Industries, Inc. Foundation
The Tappan Company
Technicon Corporation
Telex Computer Products, Inc.
The Times Mirror Company
Timex Corporation
Henry R. Towne Trust
Transamerica Corporation
Transco Energy Company
Transway International Foundation
Triangle Foundation
Union Bank
Uniroyal, Inc.
United Energy Resources, Inc.
The UPS Foundation
Warner-Lambert Company
Weyerhaeuser Company Foundation
The Williams Companies
Wilson Foods Corporation
Wilson Sporting Goods Co.
The Witco Foundation
Wm. Wrigley Jr. Company
Zapata Corporation

Appendix III

ADDITIONAL SCHOLARSHIPS INFLUENCED BY SAT OR ACT EXAMS
(partial listing)

NAME OF FUNDING ORGANIZATION	NUMBER AWARDED	TEST(S) REQUIRED	AMOUNT AWARDED
Aid Association for Lutherans	400	SAT/ACT	$ 500 to $ 7000 ea
Aid Association for Lutherans (Nursing)	25	SAT/ACT	$ 2000 to $ 7000 ea
American Postal Workers Union	25	SAT/ACT	$ 100,000 total
Continental Can Company	20	SAT	up to $ 10,000 ea
Daughters of Penelope	3	SAT/ACT	$ 2000 or $ 1400 ea
Daughters of the American Revolution	8	SAT/ACT	$ 32,000 total
Dravo Corporation	5	SAT	$ 15,000 total
Glass, Pottery, Plastics Workers Union	4	SAT	$ 36,000 total
Graphics Communications Intern'l Union	10	SAT/ACT	$ 10,000 total
Intern'l Alliance Theatrical Employees	2	SAT/ACT	$ 6,000 total
Intern'l Assoc. of Iron Workers	2	SAT/ACT	$ 12,000 total
Ladies Garment Workers Union	10	SAT	$ 90,000 total
National Eagle Scout	36	SAT/ACT	up to $3,000 ea
Pitney Bowes, Inc.	30	SAT	$ 75,000 total
Portugese Continental Union	9	SAT	$ 3,250 total
Royal Neighbors of America	22	SAT/ACT	$ 48,500 total
Service Employees International Union	11	SAT/ACT	$ 33,000 total
Sperry and Hutchinson Company	50	SAT/ACT	$ 100,000 total
Stanhome Inc.	50	SAT	$ 750 to $ 5000 ea
UNICO	3	SAT/ACT	$ 12,000 total
U.S. Air	5	SAT	$ 5,000 total
Westinghouse	75	SAT	$ 315,000 total
William C. Doherty Scholarship Program (letter carriers union)	15	SAT	$ 48,000 total

Mr. **Edwards**. Thank you very much, Ms. Rosser. We are going to have the rest of the panel testify before we ask some questions.

I might congratulate Ms. Magazine for contributing to The Morning Edition on All Things Considered. Those of us who ride in automobiles appreciate what you're doing there.

Our next witness and member of the panel is Dr. Nancy Cole. Dr. Cole is Dean of the College of Education, University of Illinois, Champaign-Urbana, IL.

Before you begin, we welcome the gentlewoman from Colorado, Mrs. Schroeder. Do you have a statement?

Mrs. **Schroeder**. No, Mr. Chairman. Thank you very much for holding these hearings.

Mr. **Edwards**. Thank you.

Dr. Cole.

STATEMENT OF NANCY S. COLE

Dr. **Cole**. Thank you.

In my various professional roles I have been a test maker, a test critic, and a student of and writer about the technical issues involved in attempting to judge whether a test is biased or not against some special group, often a group defined by race or gender.

Unfortunately, I am not able to appear before you today with simple answers and simple solutions to the very complex questions of standardized test use and race and gender differences. In fact, although my background is technical, and my work has been technical, it has led me to view the issue of standardized test use and race and gender differences as an issue that reaches far beyond the technical. In fact, the issues involved are broad social issues and at the very heart of these issues are the questions of how we view performance and opportunity differences of various sorts in this society.

When people became especially concerned with race and gender implications of standardized testing in the late 1960s, on the heels of broad civil rights concerns, the expectation of many was that we would find large artifactual effects in tests that produced the group differences that were being observed. That is, it was hoped that the group differences being observed were the fault of the tests. There were then, and there are now, bad tests, and there are bad uses of tests. But the stronger finding of a decade of study of tests and the possible bias in them has been that the differences are likely no greater in many tests than the differences all around us—in the way children are raised in their homes, in the schools they attend, and in the activities in which they engage.

There are great differences in experiences and opportunity in this country by race, socioeconomic status, and gender. Not surprisingly, these differences result in differences in performance, goals, and aspirations, also by race, socio-economic status, and gender. The bigger issue by far than the tests themselves is how, as a society, we respond to changing the experience and opportunity differences—whether we accept and resign ourselves to performance differences, or act affirmatively to try to create experiences and opportunities to overcome those differences.

Let me illustrate the complexities of the issue. Standardized tests generally show better performance by girls on school-related subjects in the elementary and middle school years. Standardized achievement tests of the school do not start with the assumption that the sexes should be performing equally at all grades through school, but set their questions based more on the curriculum in the schools and what the schools are trying to teach the children. At the early grades, the girls outperform the boys on essentially every subject.

By high school, the gap narrows and reversals in some subjects occur. There has been much discussion of the result that young women in high school as a group score more poorly on mathematics tests than do young men.

This result has raised a number of questions: are the tests biased against the young women at this stage? Are the schools biased against the young women at this stage? Are the parents biased, or are the genes biased? These questions are stated in terms of bias because many people address them that way. However, there are really far more illuminating ways to ask these questions. Some of the examples I would like to raise are the following:

Are the tests asking questions to which young men and young women have been equally exposed? We often find they are not. Should they be limited to such questions? If young men are taking advanced mathematics more frequently than young women in high school, should the high school achievement tests be limited to the types of questions and courses that the two groups are being equally exposed to? Are the questions the test asks important ones on which we care about performance?

If there are group differences, one of our very first questions to ask ourselves is, are the tests measuring something we care about? Because if they're not, then we don't care that there are group differences. But if the tests are measuring something that looks very important to us, then we had better worry about the implication of those differences.

Are the schools providing equal encouragement to young women and men to take mathematics courses? There are lots of indications that they probably are not. Do the teachers and the counselors believe in the importance of mathematics for both sexes and act on those beliefs? If there are differences in the tendencies of the sexes toward mathematics, what is being done to either reinforce those differences in the schools or to counter those differences? What should be done?

Are parents providing equal encouragement to their children of both sexes? Almost certainly not. If not, what should be done to overcome those differences?

Are young women less able to learn mathematics than young men, even after all the subtle differences of encouragement and opportunity are eliminated? Even if that were the case, should we try to counter that by looking for ways to help young women catch up, or resign ourselves to the differences?

To limit our questions to the tests and their characteristics is far too narrow a view of the issue. There are a range of questions we should ask about situations in which such differences appear. Only the first of these is: *should the test be changed*? Within this question of changing the test, we must address issues of whether the performance the test is assessing is important and relevant to the use being made of those test scores. This is the usual test validity question. In addition, we must address whether the nature of the test favors one group over another in ways that are irrelevant to the intended purpose, or whether those differences are relevant to the intended purpose. In other words, are the performance differences real ones that matter? These are the test bias questions.

Even if we judge that it is not the test that should be changed, we must ask: *should the use of the test be changed*? Here the concern is whether the use to which the test is being put does more harm than good—with the social impact of the test use. Part of this concern involves whether or not there are alternatives to this test that could accomplish the goal with less negative social impact. Sometimes there may be. However, there are instances in which the alternatives to the tests could potentially be more harmful than the tests, so one should not assume eliminating the test in favor of nontest alternatives is automatically an improvement. It might be optimistic to assume that judgments without tests for college admissions, for example, would automatically right the balance between males and females in a better way than the standardized tests do.

Part of the issue concerns whether the goal of the test use is itself socially desirable. It requires a careful weighing of social pros and cons to reach a reasonable conclusion about the total social impact of the test use in college admissions, in testing teachers, and in testing students in schools. There is a range of types of social impacts that these can have, and in my view it is not a simple question to balance the pros and cons. Finally, part of the issue is whether test users are putting too much stock in test scores or giving meaning to them which is not justified.

Whether or not we judge that the use of the test should be changed, we have the additional question: *should the experiences leading to the test performance be changed?* If the differences are important, relevant, and real, what are we as a society going to do about them for the individuals directly involved or the generations that will follow them. Concern with the tests has too often allowed us to avoid concern with this more fundamental issue. If young women are performing more poorly in mathematics at high school and college age, what should we do about it? The real need for strong, affirmative, positive action to create change is at the level of the experiences leading to test performance differences. For example, what should we do affirmatively in schools to encourage the women students to study mathematics, to help them overcome mathematics anxiety, to produce real opportunity for mathemtics learning that overcomes the variety of negative experiences with mathematics that young women have?

To point to the areas I view as even more important than the tests is not to recommend to you that the tests should be "let off the hook." We have much to learn in judging whether the performances the tests are measuring are, in fact, important and relevant to the uses made of them. We have not eliminated all possibility of irrelevant difficulty for some groups in the nature of the questions and the ways the tests are given. The issues of validity and bias are not resolved and we should continue to press the test producers toward high standards and requirements of thorough evidence to address these validity and bias issues. However, if our attention is focused only here, we may miss the even more important considerations of whether a particular type of use of even a good test is socially desirable and how we must change the different experiences of persons of different race, socioeconomic status, and gender if the goal of equal performance is to be a reasonable one.

Thank you.

[The statement of Nancy S. Cole follows:]

Hearing Date
April 23, 1987

Testimony of Nancy S. Cole
to the
Subcommittee on Civil and Constitutional Rights
for the hearing on
Standardized Test Use and Race and Gender Differences

In my various professional roles I have been a test maker, a test critic, and a student of and writer about the technical issues involved in attempting to judge whether a test is biased or not against some special group, often a group defined by race or gender. However, my expertise has not prepared me to provide neat and simple suggestions about race and gender differences on standardized tests. My learning in this area has made me very humble as it has revealed an issue of tremendous complexity and subtlety, not conducive to easy solutions that I can find. The complexities are sufficient even within the realm of technical considerations of bias and group differences. However, the issues are not and cannot be viewed as simply technical; in fact, the issues are broad social issues. At the very heart of these issues are the questions of how we view performance and opportunity differences of various sorts in this society.

When people became especially concerned with race and gender implications of standardized testing in late 1960's on the heels of broad civil rights concerns, the expectation of many was that we would find large artifactual effects in tests that produced the group differences observed. That is, it was hoped that the group differences being observed were the "fault" of the tests. There were then and are now bad tests, but the stronger finding of a decade of study of the tests and possible bias in them has been that the differences are likely no greater in many tests than they are all around us--in the way children are raised in their homes, in the schools they attend, and in the activities in which they engage.

There are great differences in experiences and opportunity in this country by race, socioeconomic status, and gender. Not surprisingly, these differences result in differences in performance, goals, and aspirations also by race, socioeconomic status, and gender. The bigger issue by far than the tests themselves is how as a society we respond to changing the experience and opportunity differences--whether we accept and resign ourselves to performance differences or act affirmatively to try to create experiences and opportunities to overcome the differences.

Let me illustrate the complexities of the issue. Standardized tests generally show better performance by girls on many school-related subjects in the elementary and middle school years. By high school, the gap narrows and reversals in some subjects occur. There has been much discussion of the result that young women in high school as a group score more poorly on mathematics tests than do young men. This result has raised a number of questions:

-- Are the tests biased?
-- Are the schools biased?
-- Are parents biased?
-- Are the genes biased?

These questions are stated in terms of bias because many people address them that way. However, there are other far more illuminating ways to ask the questions. Some examples are:

-- Are the tests asking questions to which the young men and women have been equally exposed? Should they be limited to such questions? Are the questions the test asks important ones on which we care about performance?

-- Are the schools providing equal encouragement to young women and men to take mathematics courses? Do the teachers and counselors believe in the importance of mathematics for both sexes and act on those beliefs? If there are differences in the tendencies of the sexes toward mathematics, what is being done to reinforce those tendencies or counter them? What should be done?

-- Are parents providing equal encouragement to their children of both sexes? If not, what should be done to overcome the differences?

-- Are young women less able to learn mathematics than young men even after all the subtle differences of encouragement and opportunity are eliminated? If so, should we try to counter that by looking for ways to help them catch up or resign ourselves to the differences?

To limit our questions to the tests and their characteristics is far too narrow a view of the issue. There are a range of questions we should ask about situations in which such differences appear. Only the first of these is, Should the test be changed? Within this question we must address issues of whether the performance the test is assessing is important and relevant to the use being made of those test scores. This is the usual test validity question. In addition we must address whether the nature of the test favors one group over another in ways that are irrelevant to the intended purpose. In other words, are the performance differences real ones that matter? These are the test bias questions.

Even if we judge that it is not the test that should be changed, we must ask, Should the use of the test be changed? Here the concern is whether the use to which the test is being put does more harm than good--with the social impact of the test use. Part of this concern involves whether or not there are alternatives to this test that could accomplish the goal with less negative social impact. There are instances in which the alternatives to the tests could potentially be more harmful than the tests so one should not assume eliminating the test in favor of nontest alternatives is automatically an improvement. Part of the issue concerns whether the goal of the test use is itself socially desirable. It requires a careful weighing of social pros and cons to reach a reasonable conclusion about the total social impact of test use in college admissions, for example. Finally, part of the issue is whether test users are putting too much stock in test scores or giving meaning to them which is not justified.

Whether or not we judge that the use of the test should be changed, we have the question, Should the experiences leading to the test performance be changed? If the differences are important, relevant, and real, what are we as a society going to do about them for the individuals directly involved or the generations that will follow them. The concern with the tests has too often allowed us to avoid concern with this more fundamental issue. If young women are performing more poorly in mathematics at high school and college age, what should we do about it? The real need for strong, affirmative, positive action to create change is at the level of the experiences leading to test performance differences. For example, what should we do affirmatively in schools to encourage the women students to study mathematics, to help them overcome mathematics anxiety, to produce real opportunity for mathematics learning that overcomes the variety of negative experiences young women have?

To point to the areas I view as even more important than the tests is not to recommend that the tests should be "let off the hook." We have much to learn in judging whether the performances the tests are measuring are important and relevant to the use. We have not eliminated all possibility of irrelevant difficulty for some groups in the nature of the questions and the ways the tests are given. The issues of validity and bias are not resolved and we should continue to press the test producers toward high standards and requirements of thorough evidence to address these validity and bias issues. However, if our attention is focused only here, we may miss the even more important considerations of whether a particular type of use of even a good test is socially desirable and how we must change the different experiences of persons of different race, socioeconomic status, and gender if the goal of equal performance is to be a reasonable one.

Bio on Nancy S. Cole

Nancy S. Cole is Professor of Educational Psychology and Dean of the College of Education at the University of Illinois at Urbana-Champaign. Dr. Cole received her B.A. from Rice University and her M.A. and Ph.D. in psychology from the University of North Carolina, Chapel Hill. She started her career in 1968 as a research psychologist at the American College Testing Program in Iowa City, Iowa where she later served as Director of Test Development and Assistant Vice President for Educational and Social Research. In 1975, Dr. Cole joined the faculty at the University of Pittsburgh where she was Professor of Educational Research Methodology and later Associate Dean of the School of Education. She assumed her present position in 1985.

The focal points of Cole's research and publications have been the measurement of vocational interests of young men and women, issues of bias in testing, and other general problems and issues of standardized achievement testing. She is author of a forthcoming chapter, "Bias in Test Use," in the third edition of Educational Measurement, edited by R. L. Linn.

Dr. Cole was president of the National Council of Measurement in Education in 1983 after previously serving on its board of directors. She has been Vice President for Division D of the American Educational Research Association (AERA), and Member-at-Large on the AERA Council. This spring she was named President-Elect of AERA and will assume the presidency of that organization in the spring of 1988.

Mr. **Edwards**. Thank you very much, Dr. Cole. Our next witness is Dr. Diana Pullin, associate dean, college of education, at Michigan State University, we welcome you, Dr. Pullin.

STATEMENT OF DIANA PULLIN

Dr. **Pullin**. Thank you, Congressman Edwards.

Let me begin by indicating that I come before you both as someone who is an academician who has done research on the public policy implications of the use of standardized testing, and also as someone who has served as plaintiffs attorney in a number of civil rights class action lawsuits across the country, challenging the use of standardized tests to make critical determinations about individuals. Let me also say that I share Dr. Cole's concern that some of the issues with which we need to be dealing with are issues concerning the nature of the test instruments themselves and the powerful influence those test instruments have. But, in addition, I think many of the questions we must address also concern the extent to which individuals taking the tests have had full and fair and equal educational opportunities to prepare them to compete successfully in the battlefields upon which these tests are being used.

I would also like to focus not only on the question of testing in higher education and high schools, but also on the very widespread use of testing that extends from kindergarten through grades 12 and into higher education.

As of the last time I took a count, which was late in 1984, 19 States had initiated tests to determine whether or not to award regular high school diplomas to students. Eighteen States were then, and I believe approximately 34 States are now, relying upon standardized competency tests to make determinations about initial teacher certification, entry into or exit from teacher education programs, and to determine whether veteran experienced successful teachers can retain their teaching certificates and their employment in the Nation's classrooms.

In addition, several Southern States and a number of local school districts across the country are using what might be termed "ready or not" testing to determine the eligibility of young children for entry into either kindergarten or the first grade. Promotional gates testing is being used in at least five States and numerous local school districts to determine whether students can be promoted from grade to grade. Achievement testing is used in almost every school district to make tracking or ability grouping determinations for class placement for students.

The SAT and ACT are being determined to use entry into higher education, and with increasing frequency in our latest mode of so-called educational reform, tests are often being used as the sole criterion for determining entry into the growing number of programs for gifted and talented students in this country.

Finally, State and Federal laws designed to reform the delivery of special education services in particular the Education for the Handicapped Act, which is designed to serve students with handicapped conditions has resulted in the increased use of tests to make diagnostic and placement decisions about students who are considered potential candidates for special education services.

All of these testing mandates are layered on top of, and have been layed on top in the past decade, of the very large amount of standardized testing that was already going on in our schools for the purposes of measuring State or local progress in educational achievement, for gathering nationwide data on education programs—through the National Assessment of Educational Progress—and for conducting various independent educational research, assessment, and program evaluation efforts. And, all of these standardized testing activities are applied on top of the very considerable amount of classroom testing done with teacher-made tests and with the many ready-made tests that come from book publishers in conjunction with many textbook series, particularly reading text series.

While I regard most classroom testing as relatively benign, the recent increase in standardized tests for the purpose of monitoring student and teacher accountability and for making critically important decisions about individual students and teachers is provoking growing controversy. This controversy has focused in particular upon the impact of the new testing requirements and the uses of standardized tests for minority students.

As has been known for many years, the performance of minority students on many, if not most, of these measures is often dramatically lower than that of their white counterparts. There is a concern growing among many of us that the new testing program may serve not simply as educational measurement devices but may, in addition, play a major role in redefining the nature and content of education itself and influencing the educational opportunities to which students are exposed.

While there is a good deal of information concerning these various issues, let me simply bring to your attention and highlight some information about various uses of standardized tests across the country.

For example, if we begin with the area of special education programs, for the most part both the legal and educational systems operate on the presumption that programs for students with handicapping conditions are made available to those who meet particular physical or medical criteria and who are therefore eligibile for special educational services. However, it is now quite clear that most of the determinations hinge heavily, if not exclusively, upon the use of standardized test instruments. This is at least in some part the explanation for the following kinds of demographic phenomena that are occurring.

In California, for example, in 1979, in a situation challenged in a class action lawsuit, *Larry P.* versus *Riles*, black children represented only about 10 percent of the total student population in the State of California. On the other hand, they accounted for approximately 25 percent of the enrollment in classes for students labeled as educable mentally retarded, a situation the Federal courts eventually found to violate Federal statutory principles.

If one were to look across the country, a recent analysis of data compiled by the Office for Civil Rights in the U.S. Department of Education indicates the following overall data concerning placements and average rates of placement into classes for the educable mentally retarded:

The overall rate of placement for students into those classes is about 1.50. The placement of white students is at a rate of approximately 0.87. The placement rate for Hispanics is 1.31, and the placement rate for blacks is 2.44. A similar analysis on a district-by-district basis indicates that in many large city districts, and in many southern districts, those rates are much higher in terms of the disproportions for black and Hispanic students.

On the other side of the coin, however, to look at programs for the so-called gifted and talented students, the overall placement rate is about 4.70. However, the white placement rate is 7.0, the black placement rate is 2.61, and the Hispanic placement rate is 2.57.

If one were to look with particular focus upon Georgia, a State in which it is my understanding standardized that test scores are the sole determinant of placement into classes for the gifted and talented, one finds, for example, that in the City of Atlanta's public schools, whites outrepresent blacks in gifted and talented placements by a rate of approximately 7 to 1. In Charlotte-Mecklenburg, NC, whites outrepresent blacks at a rate of approximately 11 to 1. On both sides of the coin, either at the special education end of the provision of services, or at the gifted and talented end of the provision of services, one finds dramatic overrepresentations of whites among the gifted and talented population and underrepresentation of blacks in that same population, with the reverse being the situation in special education classes for the educable mentally retarded.

The issues plays itself out again when one looks at the current wave of efforts to use teacher competency tests to determine continuation of certification or initial certification for individuals seeking to prepare for the important profession of teaching. The legal rule used under Title VII of the Civil Rights Act, of course, is that one cannot use that statute in the Federal court system to address race differences in testing unless there is a statistically significant difference between black and white rates of performance on the test.

A frequent rule used to determine disproportion is that one should look for a two-standard deviation difference in such performance. Some data I dealt with just last week indicated that in the teacher competency test being used by the State of Georgia to determine initial certification and continuing certification, there is a 119.7 standard deviation rate difference for blacks taking the examination.

There are, in short, a number of considerably troublesome questions that can be presented when one deals with this issue, and I do not envy the subcommittee having to grapple with the complexity of the issues presented here. I think, in making your determination about how to proceed in your deliberations, I would ask only that you consider very carefully not only the question of looking at the tests themselves and the extent to which we can encourage very rigorous standards for validity and reliability of the tests, the extent to which we can attempt to minimize the use of tests as the sole criterion for making decisions, and for engaging in decision making about significant issues in the lives of the student or teacher, but also to ask you to carefully consider the civil rights and educational implications of the use of these tests.

Tests are becoming more and more pervasive in the kindergarten through grade 12 culture of our schools. They are coming to be very influential in the nature of the relations between teachers and students, and they will have a growing, rather than a decreasing, level of importance in the Nation's schools and in attempts to ensure the enforcement of the civil rights of women and minorities who must work in those schools.

Thank you.

[The statement of Diana Pullin, with attachments, follow:]

EDUCATIONAL TESTING: IMPACT ON CHILDREN AT RISK

by Diana Pullin

Increased Use of Standardized Tests

A series of recent reports have accused the nation's public schools of promoting mediocrity and generated an increased interest in the use of tests to measure educational progress. Concern about the quality of public education provoked an increase in test use beginning in the mid-1970's. Since 1975, the use of tests to make critical educational decisions about students and to implement various public policy goals has increased dramatically. As of the late summer of 1984, nineteen states have initiated tests to determine the award of regular high school diplomas. Eighteen states are relying on competency tests to make determinations about teacher certification or entry into or exit from teacher training programs. Several southern states and a number of local school districts are using "ready or not" testing to determine the eligibility of young children for entry into kindergarten or the first grade. Promotional gates testing is being used in at least five states and numerous local school districts to determine grade-to-grade promotion. Achievement testing is used in almost every school district to make tracking or ability grouping determinations for class placement. Finally, state and federal laws designed to reform the delivery of special education services to students with handicapping conditions have increased the use of tests to make diagnostic and placement decisions about students (Pipho & Hadley, 1984).

These new testing mandates of the past decade add to the large amount of standardized testing already going on in our schools for the purposes of measuring state or local progress in educational achievement, gathering nationwide data on educational progress (through NAEP, the National Assessment of Educational Progress), determining access to higher education (through the SAT or ACT), and conducting various independent educational research, assessment, and program evaluation efforts. All of these standardized testing activities are applied on top of the very considerable amount of classroom testing done with teacher-made tests and the many ready-made tests that come from book publishers in conjunction with many textbook series.

Classroom teachers have long relied upon the use of tests to make assessments of individual and group progress and to gather information about the extent of individual or group educational deficiencies. Little public controversy has resulted from classroom testing for several reasons. Since teachers have available to them a wide variety of information about their students in addition to classroom tests, much of it based upon direct personal observation, the concerns about test use are minimized due to the presumption that teachers act not on the basis of test scores alone but instead use a wide array of information available to them about each student. In addition, most feel that the decisions of an individual teacher about a student do not have significant implications for the life chances of that student given the many other educators with whom the student will come into contact over the years.

Testing as a Barrier to Educational Opportunity

While classroom testing may be regarded as relatively benign, the recent increase in the use of standardized tests for the purpose of monitoring student and teacher accountability has provoked considerable controversy. This controversy has focused upon the impact of the new testing requirements on minority students, whose performance is often dramatically lower than that of their white counterparts, and a concern that the new testing programs may serve not simply as educational measurement devices but may, in addition, play a major role in redefining the nature and content of education itself (Madaus, 1980, 1985).

Educators have long known that low income, minority, and limited-English-proficient youth consistently demonstrate lower levels of proficiency on most standardized tests. These lower scores are in large part the result of the limited educational opportunities traditionally afforded the nation's low income, minority, and limited-English-proficient students. In many situations in the past, low performance on a standardized test, particularly an achievement test, would not work to disadvantage a student and could often be used to direct the student to more beneficial educational opportunities particularly targeted to the student's needs (Madaus, et al., 1980).

However, our educational history also includes substantial evidence that the otherwise relatively benign use of achievement test data to guide educational programming or planning for either individual students or groups of students can be frought with very negative unintended consequences (Oakes, 1985). For example, achievement test scores used to determine class placement have frequently resulted in the racial isolation of minority students in particular groups or tracks within school buildings. Often, these so-called ability grouping mechanisms have been adopted by schools undergoing the early years of school desegregation; here, test use has often been halted by courts on grounds that the tests were being used as a mechanism for circumventing integration (Committee on Ability Testing, 1980). While the notion of targeting instruction to students' particular educational needs is the practice advocated by almost all educators and the concept of tracking or ability grouping homogeneous students to foster such an approach sounds appealing, tracking and grouping practices have not succeeded in promoting this goal. Much available evidence indicates that, rather than helping to foster educational attainment so that students acquire more skills and knowledge and, therefore, move up and out of lower tracks or groups and into higher ones, the vast majority of lower track or group placements become dead ends. Students rarely move up onto higher level placement, in part because the diluted curriculum and instruction provided in the lower levels leaves the students enrolled there further and further behind their age peers, rather than enhancing their attainment so that they might catch up. The result of test use to determine class placement, therefore, often serves as a roadblock to access to future educational opportunities (Oakes, 1985; Labaree, 1983).

Misclassification of Minority Students

Enrollments in certain types of special education programs demonstrate a similar phenomenon. While enrollments in programs for students with handicapping conditions that are identified according to physical or medical criteria are statistically representative of the racial proportions in the population as a whole, enrollments in programs serving students with educationally-defined conditions frequently contain disproportionate enrollments of minority youth. Classes for students identified as having a moderate level of mental retardation are often populated with a disproportionate number of minority and low income youth. In California in 1979, for example, black children represented only about 10 percent of the total student population, but accounted for 25 percent of the enrollment in classes for students labeled as educable mentally retarded. Federal courts assessing the situation determined that these disproportionate enrollments resulted from over-reliance upon standardized intelligence tests to make decisions about special education placements (Committee on Ability Testing, 1983).

Given these types of trends in test performance and the use of test data, it is not surprising that new uses of tests have produced results which also include significant discrepancies between the performance of white, middle class students and the rest of the school population. While this data provokes the same types of haunting questions about what happens to minority and low income children in our schools, it also evidences a problem of even more serious magnitude. The new tests are being used to make decisions that are critical in determining the life chances of the children who take the tests. Indeed, in some instances, test scores are the only evidence considered in making significant decisions about students. When a written examination is the sole basis for determining the award of a regular high school diploma and when substantial numbers of minority students fail that test, the result is a bar to entry to the job market, to military service, and to higher education for a significant proportion of minority youth (Pullin, 1984). Further, the prospects of the loss of a diploma even after twelve years of the requisite attendance and attainment of passing grades has apparently been daunting enough to provoke an increase in the rate of school dropouts (Madaus, 1985).

While a detailed national record of results on the use of high school graduation tests is not yet available, the results from several of the states presents a vivid picture of the implications of the testing requirements for minority youth. In Florida, the first state using a test as the criterion for determining the award of a high school diploma, the initial scores on the test indicated that black students failed the test at a rate ten times greater than the failure rate for whites; statistical analyses indicated that these disproportions were more highly correlated with race than student socioeconomic status, a factor often used by educators to explain school performance. The first time the requirement was actually imposed after court orders had placed a four year moratorium on the use of the tests due to its unlawful effects, 57 percent of the students who failed to meet the test requirement were black in a student population that was only 20 percent black (Madaus, 1983).

Testing and the Handicapped

Another population that may be particularly disadvantaged by the use of tests as high school graduation requirements are students with handicapping conditions. Many times in the past, students with handicapping conditions were able to attain high school diplomas, enter into the world of work, and attain a degree of economic self-sufficiency. This same category of students now is barred from diplomas, often because the nature of handicapping conditions is such that success on a paper and pencil test is impossible (Pullin, 1984). Particularly when passage of a graduation test is supposed to represent so-called "real world competency," the result may be cruel for many students who fail the tests are able to function competently in the real world. Evidence such as this raises real questions about what it is that the tests are measuring and the accuracy with which the measurements occur. The lay public, as well as many educators, place the same faith in standardized, paper and pencil tests that they place in the thermometers with which they read their body temperature. A thermometer reading of 99.1 degrees may not necessarily mean that you have a fever; thermometers are not perfectly accurate and the usual body temperature of each of us varies such that the "normal" temperature of 98.6 is simply an approximation. Education tests work in a way quite similar to thermometers. Our nation has recently become quite enamored with the use of tests as a means of measuring the success of the educational process. Like thermometers, tests are not always accurate and some are much less accurate than others. One large difference between thermometers and tests, however, is that most of us would agree that knowing your body temperature when you are not feeling well may be a useful piece of information. The same cannot necessarily be said for educational tests. Thinking that something is wrong with our schools, or with particular children in our schools, because of poor test performance neither tells us with accuracy that something is wrong, what is wrong, or why something is wrong. When Johnny can't read, his parents and teachers ordinarily know he can't read and don't need a test score to confirm this fact. A more troublesome fact, and one which tests rarely help, is that too often we don't help Johnny learn to read or to read better. Tests have little to do with these problems, particularly the new wave of tests now popular with policy makers; these tests provide no diagnostic information about why Johnny can't read and are therefore of no help in correcting the situation.

Limitations of Testing

The importance of understanding what it is that tests can and cannot tell us is critical. Not all tests are accurate measures of the skills and knowledge they purport to measure and even the more accurate tests are at best approximations. In addition, the public is often led to make generalizations about individuals that cannot be supported on the basis of test information. Teacher competency tests provide a good example of the types of misconceptions that tests can generate. While those working closely with a testing program may be well aware of the content covered by the test and the permissable inferences that can be drawn from test scores, the public use of test information may be far broader and less defensible. For example, the public might quite understandably feel that a teacher or a candidate for a teaching certificate who fails a teacher competency test is an incompetent teacher. However, despite what the names of these tests imply, none of the so-called teacher competency tests in any way measure the on-the-job performance of teachers and none of the tests tell us whether teachers can in fact teach and teach well. This may explain why in one state two of the teachers who failed a teacher competency test had recently been named teacher of the year in their respective school districts.

Although the new tests are of little help in improving the education of individuals, they may exert a powerful influence upon what happens in schools. Research on the effects of student competency tests both here and in Europe indicates that the content covered on the tests tends over time to control the skills and knowledge covered in the curriculum. This result has led several commentators to charge that the long-term effect of the use of minimum competency tests will be the further dilution of the content of instruction such that the minimums covered on the test become the maximum limits for instruction (Madaus & McDonough, 1979). This charge is particularly disconcerting given the growing body of evidence that, while the proficiency of students in basic skills areas has been increasing over a period of years (predating even the introduction of most of the student competency tests), student competence in the higher order skills of complex and critical thinking and problem-solving has been declining (Madaus, 1981). Given the power of externally imposed tests to influence the content of instruction and the current focus of the tests upon minimum basic skills, the tests may exacerbate the growing problem of declining achievement in higher order skills.

The Politics of Testing

Two more general problems inherent in the new testing movement provide clear warning that the tests will probably not promote the types of educational reform and accountability test proponents are seeking. First, it is well to remember that, for the most part, the new testing mandates are the result of political reforms, not educational reforms (Wise, 1978; Madaus, 1985). The new programs have most often been imposed by legislators and state and local school board members. The programs are not the result of efforts to apply state-of-the-art insights from educational research into practice. Indeed, research in this area suggests that the new programs may well have more deleterious consequences, such as diluting curriculum content, than the wide-ranging benefits test proponents have predicted.

Second, we have only begun to understand the implications and impact of these programs. Test proponents tout dramatic increases in test scores, particularly for minority students. However, with insufficient data on what it is that the tests are measuring, we cannot know if students now, in fact, know more or if, instead, they have enhanced their test-taking skills, reduced their test anxiety, or been given tests written at a less difficult level. Further, because most states fail to maintain adequate or accurate data on student dropouts, we have no way of knowing if pass rates have gone up on part because the proportion of discouraged students taking the test has gone down (Madaus, 1981).

Tests are appealing. They appear to afford the public an easy, even scientific way of measuring the progress of our educational system. However, while the public has come to believe that improved test scores represent educational progress, test scores are only surrogate measures of real learning, the acquisition of important, useful skills and knowledge. To that extent, the new tests may appear to demonstrate that our students possess "the right stuff" while, instead, all that we have achieved is an illusion of educational progress, a portrait sketched at the expense of many youngsters who are disadvantaged by these testing schemes.

BIBLIOGRAPHY

Cawelti, G. (1978, May). National Competency Testing: A Bogus Solution. Phi Delta Kappan, 619-621.

Committee on Ability Testing. (1982). Ability Testing: Uses, Consequences, and Controversies. (Part I). Washington, DC: National Research Council.

Down and Out in the Classroom: Surviving Minimum Competency. (1979, January). Principal, 58, 12-59.

Gallagher, J. & Ramsbotham, A. (1978, October). Developing North Carolina's Competency Testing Program. School Law Bulletin, IX, 8-14.

Gould, S. (1981). The Mismeasure of Man. NY: W.W. Norton & Company.

Haney, W. & Madaus, G. (1978, November). Making Sense of the Competency Testing Movement. Harvard Educational Review, 48, 462-84.

Houts, P. (1977). The Myth of Measurability. New York: Hart Publishing Company, Inc.

Hyman, R. (1984, March/April). Testing for Teacher Competence: The Logic, The Law, and The Implications. Journal of Teacher Education, XXXV, 14-18.

Labaree, D. (1983). Setting the Standard: The Characteristics and Consequences of Alternative Student Promotional Policies. Philadelphia: Promotion Standards Committee of Citizens Committee on Public Education.

Levin, H. (1978). Educational Performance Standards: Image or Substance? Journal of Educational Measurement, 15, 309-319.

Madaus, G. (1981, October). NIE Clarification Hearing: The Negative Team's Case. Phi Delta Kappan, 63, 92-94.

Madaus, G. (1983). The Courts, Validity, and Minimum Competency Testing. Boston: Kluwer-Nijoff Publishing.

Madaus, G. (1985, May). Test Scores as Administrative Mechanisms in Educational Policy. Phi Delta Kappan, 611-617.

Madaus, G. & Airasian, P. (1977). Issues in Evaluating Student Outcomes in Competency-Based Graduation Programs. Journal of Research & Development in Education, 10, 79-91.

Madaus, G., Airasian, P. & Kellaghan, T. (1980). School Effectiveness: A Reassessment of the Evidence. New York: McGraw-Hill Book Company.

Madaus, G. & McDonagh, J. (1979, June). Minimal Competency Testing: Unexamined Assumptions and Unexplored Negative Outcomes. Paper presented at the annual conference on Large- Scale Assessment sponsored by National Assessment of Educational Progress, Denver, CO.

McClung, M. (1979). Competency Testing Programs: Legal and Educational Issues. Forham Law Review, 47, 698-701.

Nathan, J. & Jennings, W. (1978, May). Educational Bait-and-Switch. Phi Delta Kappan, 621-625.

Oakes, J. (1985). Keeping Track. Connecticut: Yale University.

O'Hare, W.P. (1979, October). Race, Socioeconomic Status, and Competency Testing. (Research Paper Series). National Social Science and Law Project, Washington, DC.

Perrone, V. (1979). Competency Testing: A Social and Historical Perspective. Educational Horizons, 3-8.

Pipho, C. & Hadley, C. (1984, July). State Activity Minimal Competency Testing. Clearinghouse Notes.

Pullin, D. (1984). "Minimum Competency Testing: A Review of the Case Law." School Law Update, Chapter 3, 161-174.

Pullin, D., Sedlak, M. & Wheeler, C. (1985). Proposals for Raising Academic Standards in Secondary Schools: Will the Public Get What It Wants?

Smith, G.P. (1984). The Impact of Competency Tests on Teacher Education: Ethical and Legal Issues in Selecting and Certifying Teachers. In M. Haberman's Research in Teacher Education.

Spady, W. (1977, January). Competency Based Education: A Bandwagon in Search of a Definition. Educational Researcher, 6, 9-14.

Wise, A. (1978, May). Minimum Competency Testing: Another Case of Hyper-Rationalization. Phi Delta Kappan, 596-608.

ABOUT THE AUTHOR

Diana Pullin holds both a Ph.D. in Education and a law degree from the University of Iowa. For the past ten years, she has been actively involved in testing issues as an education researcher, an educator, and a litigation attorney. She has represented parents and students, school districts, and teacher associations in disputes involving school testing, educational accountability and educational equity for minority and special needs students. She is most known for her representation of the students and parents who brought the landmark federal court challenge to Florida'a minimum competency testing program.

At the present time, Dr. Pullin is Associate Dean of the College of Education at Michigan State University.

CONTACTS ON BACKGROUNDER #2

Dr. Pamela George
1025 Lakewood Avenue
Durham, NC 27707
919-489-0296

Dr. George Madaus
Director, Center for the
Study of Testing, Evaluation
and Educational Policy
Boston College
Chestnut Hill, MA 02167
617-552-4521

Dr. Vito Perrone
Center for Teaching and Learning
University of North Dakota
Grand Forks, ND 58202
701-777-2674

1984-85 Average ACT Composite Scores*

18.6	Overall
19.4	White
15.9	Puerto Rican, Cuban, Other Hispanic
19.1	Asian American, Pacific Islanders
14.6	Mexican American, Chicano
13.9	American Indian, Alaskan Native
12.5	Afro-American, Black

*from: ACT Issue Gram #6 (January 1986)

Impact of the ACT In Mississippi**

ACT Scores for Three Historically Black Universities in Mississippi
Mean of Entering Freshmen, 1985-86

Alcorn State	13.07
Jackson State	14.01
Mississippi Valley State	12.53

ACT Scores for Other Mississippi Universities

Delta State	19.06
Mississippi State	21.19
Mississippi University for Women	20.08
University of Mississippi	20.83
University of Southern Mississippi	19.61

**from ACT Report to Board of Trustees of State Institutions of Higher Learning in Mississippi

The Condition of Education

1986
Edition

Statistical Report
Center for Education Statistics

Edited by:
Joyce D. Stern and Mary Frase Williams

U.S. Department of Education
William J. Bennett, Secretary

Office of Educational Research and Improvement
Chester E. Finn, Jr., Assistant Secretary

Center for Education Statistics
Emerson J. Elliott, Director

A. Outcomes: Transitions

High school completion by race and ethnicity

In examining the outcomes of our schools, one important measure is whether students are able to complete the educational process. If they do not finish high school, then it is doubtful that they have obtained sufficient knowledge, skills, and abilities many citizens believe necessary to function productively in society.

Thus, one outcome measure of education is the extent to which students complete high school with classmates about the same age. The data in the accompanying table reflect percentages of students who have successfully completed 12th grade or the equivalent at ages 18-19, and ages 20-24.

The public generally expects 18- to 19-year-olds to have a high school diploma. And, indeed, most do. However, as can be seen from the table, many students take a longer period of time to complete their high school education. For example, the percentage of 20- to 24-year-olds having obtained a high school diploma or its equivalent is about 10 percentage points greater than that for 18- to 19-year-olds.

The data have been computed from tabulations from the Bureau of the Census Current Population Surveys. These data are collected from household interviews and include information on individuals who have completed 12 or more years of schooling or who have obtained an alternative credential such as a General Educational Development (GED) certificate.

Table 1:8

High school completion by race and Hispanic origin, persons ages 18 to 19 and 20 to 24: 1974 to 1985

	Age 18 to 19				Age 20 to 24			
Year	Total	White	Black	Hispanic[1]	Total	White	Black	Hispanic[1]
	Percentage of age group				Percentage of age group			
1974	73.4	76.2	55.8	48.9	83.9	85.6	72.5	59.0
1975	73.7	77.0	52.8	50.0	83.9	85.9	70.5	61.3
1976	73.1	75.4	58.2	50.9	83.7	85.4	71.9	58.0
1977	72.9	75.7	54.9	50.7	83.7	85.1	73.4	56.6
1978	73.5	76.3	54.9	48.9	83.7	85.2	73.5	58.7
1979	72.8	75.3	56.4	53.7	83.2	84.9	71.8	55.8
1980	73.7	76.1	59.3	46.1	83.8	85.1	74.3	57.1
1981	72.5	74.8	59.6	47.2	83.7	85.0	75.7	59.3
1982	72.0	74.5	58.2	51.7	84.1	85.4	76.2	60.2
1983	72.7	75.6	59.1	50.3	83.3	84.6	75.8	56.6
1984	73.3	75.5	63.0	58.3	84.6	85.7	79.3	60.7
1985	74.6	76.7	62.7	49.8	85.3	86.0	80.8	67.3

[1] Most of the year-to-year differences in completion rates for Hispanics are not statistically significant due to the small size of the Hispanic sample

NOTE: Asians are not included in the analysis because they are not identifiable from the October Current Population Survey data tapes

SOURCES: U.S. Department of Commerce, Bureau of the Census, Current Population Reports, Population Characteristics, Series P-20 *School Enrollment—Social and Economic Characteristics of Students: October* (various years), Current Population Surveys (unpublished tabulations)

CHART 1:8A -- **High school completion rates by race and Hispanic origin, persons age 18-19**

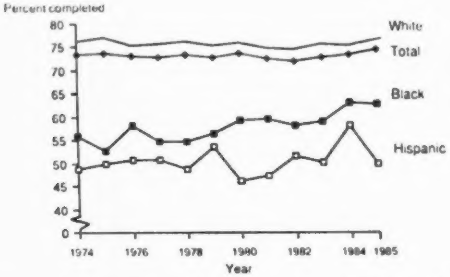
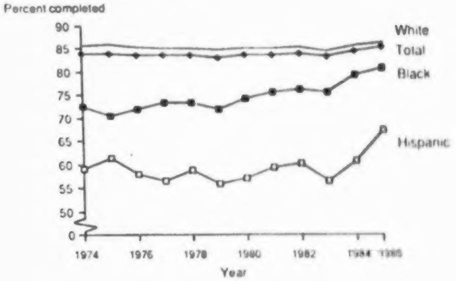


CHART 1:8B -- **High school completion rates by race and Hispanic origin, persons age 20-24**



SOURCE: Bureau of the Census, Current Population Reports, Series P-20

- Nationally, slightly less than three-quarters of all 18- and 19-year-olds have completed high school.
- The proportion of 20-to 24-year-olds who have completed high school has held steady at about 84 percent since 1974.
- The high school completion rate among blacks, for both 18-to 19- and 20-to 24-year-olds, has increased in the last decade. The rates for both blacks and Hispanics still lag far behind those of whites.

C. Context: Student Characteristics

Participation rates for higher education by race/ethnicity

Americans have prided themselves on having one of the most democratic systems of education in the world. The goal of equal access for all qualified youth has long been held as a major objective of our educational system. A measure of the national progress toward that goal is participation rates in higher education of various populations. This indicator looks at participation rates of whites, blacks, and Hispanics aged 18-24 since 1970.

Black participation rates improved dramatically from 1973 to 1976. Hispanics also increased their participation between 1972 and 1975 although they remained somewhat lower than the white rate. [illegible] participation rates have increased since 1979. Black participation rates declined in the late 1970's and have been relatively stable since then. Year-to-year differences for Hispanics since 1975, however, are not statistically significant.

Caution should be used in interpreting the data presented here. The racial/ethnic definitions the Bureau of the Census uses are not mutually exclusive. Therefore, direct comparisons between Hispanics and whites or blacks are not possible. Whites and blacks are defined as racial groups, whereas Hispanics are defined as an ethnic group and can be of any race.

Table 2:9

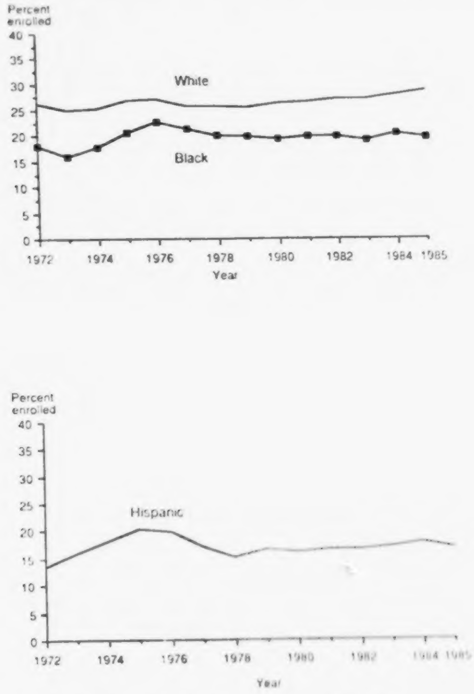
Participation rates of 18- to 24-year-olds in higher education by race/ethnicity: 1970 to 1985

Year	Racial/ethnic group		
	White	Black	Hispanic
	Percent enrolled		
1970	27 1	18 7	—
1971	27 2	18 2	—
1972	26 4	18 1	13 4
1973	25 0	16 0	16 0
1974	25 2	17 9	18 1
1975	26 9	20 7	20 4
1976	27 1	22 6	19 9
1977	26 5	21 3	17 2
1978	25 7	20 1	15 2
1979	25 6	19 8	16 6
1980	26 4	19 4	16 1
1981	26 7	19 9	16 7
1982	27 2	19 8	16 8
1983	27 0	19 2	17 2
1984	28 0	20 4	17 9
1985	28 7	19 7	16 9

—Not available

SOURCE U S Department of Commerce Bureau of the Census Current Population Reports *School Enrollments—Social and Economic Characteristics of Students* October (Series P-20), various years

CHART 2:9 -- **Higher education enrollment rates of 18- to 24-year-olds by race / ethnicity**



SOURCE Bureau of the Census, Current Population Reports, P-20

- Participation rates for minorities increased during the early 1970's.
- The proportion of blacks 18 to 24 years old attending postsecondary institutions increased after 1973, declined after 1976, and has been relatively stable since 1978.

Mr. **Edwards**. Our thanks to all the members of the panel. It was very splendid testimony.

The gentlewoman from Colorado, do you have any questions?

Mrs. **Schroeder**. Thank you, Mr. Chairman. I thank all the witnesses also for their testimony.

I guess the question I have is—what I think I hear you saying—is that while young women score lower on the tests, and the tests are supposedly to be predictors of performance in college, that when they get to college they do better. Therefore, the test is really not valid and we're not just whining about the fact that we haven't had the same background or the same math classes. It really is not predicting what women do once they get to college.

Is that correct? Is that the bottom line?

Ms. **Rosser**. Yes, it is correct. It is not predicting. That is the reason that these college entrance exams are given, that they are supposed to predict future performance. They are not doing that for girls. They are also not valid in respect to their past performance in high school, because girls are getting better grades.

Mrs. **Schroeder**. As a mother of a 16-year-old daughter, and listening to all of them now taking their tests and coming home, it really is very interesting because I see that going on. You see how perplexed some of them get if they're not scoring the way they thought they should be scoring. They are beginning to wonder if their high school performance wasn't valid, or if they had charmed their high school teachers, if maybe suddenly they're not as good as they used to be. They really start having incredible self-doubts about that.

But I guess my frustration is, if the test isn't adequately predicting what women do when they get to college, and everybody can see that, why in the world don't they change the test? That was the whole purpose for the test and I don't understand why universities haven't changed it, or why they are relying on the tests so much, if that is true.

You know and I know that high school counselors tell these kids, "Hey, you don't score here, you don't apply; you don't score here, you don't apply." I mean, it is really the key to the college door and everything is tied to that score. Forget what they did in high school as far as grades, or whether they were taking college level courses; none of that matters. They really hang so much of it on the books you buy at the bookstore, or in the way that the college counselor directs you on that score. So a lot of young girls are beginning to think that maybe they were a fraud, you know, that something has happened.

So why haven't they changed it? One of you has been suing, and others have been preaching. Why do the colleges insist on continuing to use them if they don't predict, and why won't the test people change it?

Ms. **Rosser**. Well, I don't know why the test people won't change it exactly. I've had a lot of discussions about that with them. They say that they feel the girls just aren't taking enough math and science, and if they would just take more, they would do better. But, in fact, girls have been taking more math and science and doing worse. The gap has been widening.

The colleges use these, it is said, because that way you do get a larger male student body, if you use these test scores, because the boys' scores are higher. So if you want to keep your student body more male than female, a very good way to do that is to use these tests.

But I am very upset about the fact that girls work very hard in high school and college and are not being rewarded by having the same opportunities to go to prestigious schools, to go to research universities, and that their whole life, their whole work up to that point, can be just downgraded by one number.

Mrs. SCHROEDER. My understanding is that the difference in the math gap between boys and girls is really very small, that the number of boys who take four years of math versus the number of girls who take four years of math, and then take the test, is fairly small. In addition, the test supposedly only goes to geometry, which is three years of math in most high schools. Therefore, the math gap really shouldn't make that much difference anyway.

Ms. ROSSER. Actually, the College Board data says that 50.5 percent of girls, versus 57 percent of boys, take four years of math.

Mrs. SCHROEDER. So it is very small.

Ms. ROSSER. Yes, it is very small. Also, it is true that geometry is supposed to be the most advanced math you need to take for this test. All college-bound girls that I know of certainly take geometry in their sophomore and junior year. So I really don't understand why this math gap is happening, either.

Dr. COLE. Could I respond to those comments, also?

Mrs. SCHROEDER. Sure.

Dr. COLE. I think we have misrepresented the situation if we leave with the impression that the college admissions tests don't predict college performance for women but do for men. That is just not the case. In fact, the data shows that women are the group that college admissions tests predict best for of any group, in terms of the relative relationship between the test scores and the college grades.

The prediction phenomenon that Ms. Rosser is referring to is the question of the statistical relationship of the level of performance to the test score that's based on complex statistical regression procedures. The sorts of differences referred to are very small. In fact, it is not clear from many perspectives whether they mean the interpretation she's giving to them at all. But to be left with the notion that the tests don't predict for women and do for men is just not a correct notion.

Mrs. SCHROEDER. So you would say the test doesn't predict for either one?

Dr. COLE. No, I would say the test predicts some for both. It certainly doesn't tell the whole story, but it's——

Mrs. SCHROEDER. Well, if it doesn't tell the whole story, then why do they rely so heavily on it? I sit there listening to young kids and watching all this counseling going on. Let me tell you, the score is 90 percent of what everbody focuses on. If it's not a good predictor, then why don't they change the tests across the board to be a better predictor for men and women?

Dr. COLE. You see, the tests are almost as good a predictor as the high school record. If you are dealing with a situation of very competitive admissions, with large numbers of applicants, the best information you can find to decide who are the students most likely to succeed are, first, the high school record of the student, and second the standardized test scores. Both are used.

I would not want to sit here and defend the situation in which an institution uses only the test score and not the high school record. My institution puts very heavy weight on the high school record as well as weight on the test scores. It worries me very much that we don't spend time reading what these kids write and judging letters of recommendation about the students. But the volume and numbers involved often preclude that.

I think, in fact, the publicity often oversells the role of the test in college admissions because it is something easy to focus on and easy to grab hold of. I think we are misled, to some extent, in the public in terms of how important the test scores are in that actual decision.

Mrs. **Schroeder**. I really don't know. My job, before I did this, was I used to analyze tests for jobs for the State of Colorado. I remember going through the ones for pilots and finding out how it was very culturally biased. They would have timed tests. I remember one in particular where they would show you pictures of a living room and you had so many minutes or so many seconds to figure out what was wrong. If you had a cracked window and you came from a neighborhood where cracked windows were normal, I mean, who's going to scratch the box? We could find things like that all the time, especially going to low income and to females, that had nothing to do with whether or not you could fly the airplane. If you were testing for whether or not you could fly the airplane, then I would have no problem.

We did the same thing with the Foreign Service exams when I first got to Congress. Then I was really able to get into them. I found that the Foreign Service exams that the U. S. Government used discriminated against every group in America except white males from four universities. Beyond that, who cares who won the Cannes Film Festival in 1952, and what does that have to do with whether or not you know how to administer a foreign aid program? Nevertheless, that's what we tested for. It had no applicability under what we require in the employment area, which is the old *Duke Power* decision—you know, where you can't test a janitor for his knowledge on classical music because, as a janitor, he doesn't have to know about classical music. You can test him on wax.

But I have always been very disturbed that I haven't seen that kind of sensitivity among colleges, and I think it is much tougher now because it's becoming much more elitist now, with the tremendous cost of college. I think a lot of parents are saying to young people, "If you can get into a top school, terrific; if you can't, we're not going to pay the difference."

I don't think this is *de minimis*. I think this is much more critical because it is channeling which way kids go now, just because of the phenomenal costs. I gave my kids T-shirts for Christmas saying "This kid is his mom's Mercedes." It's true. But for those kids, you could drive all sorts of cars. A lot of families don't have that option, and if they can't get into a top school, they aren't going to pay the money.

As I listened to the counselors, it is not much different than when I went to school. They are hanging the whole thing on the score. It may be wrong, the schools may not mean it, but that's what they are doing, and I really salute those of you who are trying to get this changed. If it discriminates against men, too, then it's wrong. But young people's whole lives are being changed by these test scores—black, white, Native American, male, female—and I think we ought to do everything we can to make them as accurate a predictor as possible, as we have done in the employment area. I think education is behind the curve on that.

Thank you, Mr. Chairman.

Mr. **Edwards**. Ms. LeRoy.

Ms. **LeRoy**. Dr. Cole, to follow up on the discussion that you and Mrs. Schroeder were having, I agree that you can't look at test results in isolation and just focus all of one's energies on these tests. You suggested that perhaps they are being missold or that the publicity surrounding them creates the perception of greater emphasis on these tests than is really being placed on them or ought to be placed on them. But the fact is that that perception is there, and I think it's more than a perception.

For example, the Secretary of Education has these famous wall charts that we've all seen—and I can't unfold it; it's too big—which basically evaluate the States on their educational performance based on SAT scores. So that entire school districts perceive themselves, and obviously, individual teachers, perceive themselves based on their students' performance on these tests. And I assume teachers are evaluated on them and individual schools are evaluated on them.

What can be done to get around that kind of problem when the chief educator in the country is, in fact, contributing to the problem?

Dr. **Cole**. When you ask that question, we move from the issue of what's wrong with the test to what's wrong with the use of the test. Every issue that Dr. Pullin raised, for example, is a question of social policy with respect to the use of the test. The Secretary's wall chart is an issue of the social policy implications of the use of the test, and the wall chart is an abysmal use of the SAT. It's almost a ridiculous use of the SAT scores.

Most of the concerns that Dr. Pullin raised are concerns about the overuse of tests, the use of tests that I think can have inappropriate social consequences that I'm very much concerned about. That is still a different question than the question of how we should change the SAT, if we should. It's a question of whether we should use it in certain ways, even if we had it perfect, the way we wanted it. It still would not work for the Secretary's wall chart.

Ms. **LeRoy**. Well, I don't want to look at those issues in isolation because I think they go together to create a problem for education in this country, with respect particularly to women and minorities, but also education generally.

I would ask you what could be done to reduce bias in the test, assuming that it's there, but I also want to ask you what can be done to deflate some of this overemphasis on or misuse of these types of tests. I realize that may be a Ph.D. thesis here.

Dr. Cole. Well, that's a very hard question for me to answer. In fact, in answering it, I think I would come back to trying to understand why we came to such a wide use of tests. It is my view that we came to it because we were unwilling to impose standards and make difficult professional judgments in other ways, in better ways, in our educational system. We came to use standardized tests for admissions to college more and more because we were more and more suspicious of the grades that kids get in high school and the quality of those courses the students are in.

We use tests for graduation from high school because we're suspicious that the educational system hasn't set standards for itself internally to use the better information. We use tests for promotion for the same reason. We use tests for identifying kids for special education classes, at either end of the scale, because we haven't trusted the professional judgment of the educational system.

I think, fundamentally, we are not going to be able to thwart this overuse of tests until we make some serious changes in the quality of the educational system and the standards that we have internal to the system, where we can use better information than just external standardized tests to support some of these decisions.

But that is not an easy solution. You see, one of my dilemmas in answering your question is what would be happening to college admissions without the test? What would be happening in putting kids in gifted programs without the tests? I am concerned that it could even be worse. I am concerned that, without the tests, at least to identify some of the bright, black kids that ought to be in those gifted programs, we could have even more of them being excluded.

Mrs. Schroeder. Would counsel yield on that point?

Ms. LeRoy. Yes.

Mrs. Schroeder. What I don't understand, though—I don't think we're arguing to do without the tests. The question is, there is a lot of reliance on the tests and we understand that we would like to change a lot of other things in the public and private education system so that schools could rely on that more. But why not make sure the test is as fair as it can be, then?

Dr. Cole. Well, we certainly should do that. But you have given examples of the worst in tests, and there are other counter-examples of tests that I don't know how to change to make more fair. That's our dilemma.

Mrs. Schroeder. Right. But let's change the worst and try and make it more gender-neutral and minority-neutral.

Dr. Cole. Absolutely.

Mrs. Schroeder. I think we're really in agreement.

Ms. LeRoy. Let me ask the other two witnesses the same general question, and that is, what can public policymakers, people in Congress or people at the Department of Education, what should they be doing, or can anything be done to assure both test equity and validity and proper use of the tests?

Ms. Rosser. I feel these tests should be predictive. They should predict what they're supposed to. Unless we feel that grades are completely erroneous, which I don't think anybody does, I think that the tests should correlate with the grades that students get, both for women and for minorities, however that is done. I feel that is what Congress should be looking into and requiring of test publishers.

Ms. LeRoy. Dr. Pullin.

Dr. Pullin. I think, as Dr. Cole suggested, this is a complicated issue. I think the testing industry could do far more than has been done to alleviate some of the problems of unfairness.

I would ask Congresswoman Schroeder to think about what she means when she talks about her goal of fairness and to be more explicit about it, because there is a good deal on those tests that is, in fact, representative of a culture and reflective of a culture, the culture of schools, and it is predominantly a white male culture. To make it look more white-maleish is not what I think any of the three of us are talking about.

When I think about the kinds of things that this body is able to do to address the kinds of issues we're talking about, they are, for the most part, the kinds of legalistic approaches that have been used to some extent successfully in the past. As the Congresswoman noted, we have made some fairly substantial gains in the employment testing arena. There is some discussion that those gains will be lost because the EECC Guidelines are under discussion for considerable dilution in terms of validity and reliability requirements. I think that would be a terrible disservice to the kinds of populations we're concerned about here, to allow those to be diluted.

Similar kinds of more rigorous standards need to be employed in the educational testing arena. To some extent they have been. For example, if you look at the regulations under the Education of the Handicapped Act, although these are not widely enforced, there are very specific regulations talking about lack of bias and talking about use of multiple criteria information to make determinations about individuals. Those kinds of standards are available. They are not available in every arena, among the educational arenas that we're talking about, and they are not being enforced.

Ms. LeRoy. Thank you.

Mr. Edwards. Does minority counsel have any questions?

Mr. Slobodin. Thank you, Mr. Chairman.

Good morning. I wanted to first ask Miss Rosser and Dr. Pullin, do you currently have, or have you ever had, any affiliations with the Fair Test organization?

Ms. Rosser. I have, for the last three months, been a consultant to Fair Test. That consultancy actually ends today.

Mr. Slobodin. OK.

Dr. Pullin?

Dr. Pullin. I know the people at Fair Test and I have talked to them about these issues in the past.

Mr. Slobodin. Let me talk to Miss Rosser for a moment about LSAT's. Reading from a passage in a book by Cynthia Fuchs Epstein, called "Women in the Law", she writes here—she is making the point that "the problem raised by preference for women is unlike the problem of other minority group preferences because women applicants have generally been better qualified than men." Then she proceeds to support that proposition, that "the average law school admission test score did not vary significantly by sex."

The law school admission test counselor reported in a study of LSAT scores for 1973-74 that the mean test score for both men and women was 527. Of registrants for the LSAT in 1973, 75.2 percent were men, 24.8 percent were women. In 1974, 75 LSAT surveys revealed that women had a slight edge over men in law school admission test scores. The mean test score for male registrants was 522, while the same score for women was 524.

It says that "women have done well in LSAT's by the standards set for law students." A 1972 study commissioned by the law school admissions counsel determined bias of sex and the tests showed that 1,150 males used as a comparison group, with 1,165 females, scored approximately 10 points lower and had a mean writing ability score approximately 7 points lower than women. Women did better on four of the six sections in the LSAT—reading comprehension, reading recall, error recognition and sentence correction. Men did better on one section, data interpretation. The two groups scored about equally on the principles and cases section.

What is the story here?

Ms. **Rosser**. Well, I am not an authority on the LSAT, but I have talked to people who have done research on this. They say that the women who take this test are probably about 10 times better verbally than the men and, in fact, they should be doing even better on the test than men than they are doing.

Mr. **Slobodin**. Yes, but in your testimony you mention as an example of bias in testing the fact that, in reading comprehension questions——

Ms. **Rosser**. I was talking about the SAT there.

Mr. **Slobodin** [continuing]. SAT's. And the women score higher on reading comprehension. That's where you're pointing out where the bias is, but that's where women are scoring higher.

My question to you is, let's talk in terms where there has been a disparity, and that's in the math area.

First of all, what is a bias? I mean, if there's a one point difference between the sexes, would you consider that bias? How about five points? Of what threshold are we talking about bias?

Ms. **Rosser**. I think when it has a major impact on people's lives, that's negative, that is bias to me. I think it's having a major impact on women's lives. That, to me, is bias.

Mr. **Slobodin**. But what is the impact? Where are women not getting into—Are you saying they're not getting into Harvard? Let's name some schools here.

Ms. **Rosser**. All right. Fewer women get into Harvard than men. Fewer women are getting into all the Ivy League schools than men. Fewer women get into the other prestigious schools than men. There is actually national data on that.

Mr. **Slobodin**. How do you explain this rise, then, in women coming into the law schools? In fact, you say at the beginning of your testimony here, "What struck me first when I looked at these tests was the overwhelming number of males that populated them—all of whom were engaged in traditional occupations like doctor and lawyer * * *" What has happened in the last 10 years has been phenomenal growth.

Wouldn't you concede that there has been a phenomenal change in the legal profession? We ought to be taking a look at that and saying, "Well, it's working in the legal profession. We have women now taking an interest and becoming lawyers. The tests aren't stopping that growth." We ought to be looking at ways of extrapolating that for math and science. How does it follow that we need to change the test?

Ms. ROSSER. Well, yes, there has been a tremendous growth in the legal profession. It's about time we had more women getting into the legal profession.

One of the things about the LSAT was that there was a lot of math on it. It didn't relate at all to being a lawyer. And because of a certain amount of testing reform pressure, some of that math was taken off.

Mr. SLOBODIN. When was that?

Ms. ROSSER. Oh, within the last five years, I believe.

Mr. SLOBODIN. Yes, but this was before. I'm citing statistics before they took the math out.

Ms. ROSSER. But some of those statistics also show that women did less well, and also that they——

Mr. SLOBODIN. Yes, but they still scored as well, if not better, than men on those tests.

Ms. ROSSER. But they did less well on the data.

Mr. SLOBODIN. So what? They did better than men in all the other sections. When you combine them, they actually had a slight edge.

Let me go on to an article which publicized the report you released last week in the New York Times. The reporter writes here in the article that your study "offers no analysis of standardized tests and gives no examples of biased questions. The findings are based on the conclusion that, because girls earn better grades than boys in high school and college, they should do as well or better on the tests."

Now, I would like to discuss this premise. Have you studied whether or not there is any bias in grading in high school courses? Why should we consider that more reliable than a question that asks "what's the circumference of this cup"?

Ms. ROSSER. I think that girls, historically, over the years have been getting better grades in high school and college, and they have been doing less and less well on these tests. I think that is bias. I don't think that I have to come up with specific questions that are biased. I think this is something that the people who know about tests will come up with. ETS knows which questions they are. I think we should really look at the effect this is having on people's lives, and that is a bias effect.

Mr. SLOBODIN. You don't see the potential for bias in—What about Dr. Cole's point, that we could have the potential for a lot more bias without the use of these tests?

Ms. ROSSER. Well, I think that society is biased against women. There's no question about that. And I think they are doing quite a good job of overcoming this handicap in the classroom.

Mr. SLOBODIN. Let's talk about the Educational Testing Service. We are going to get some testimony from Dr. Dwyer, where she says about 80 out of 125 people that are involved in developing these standardized tests are women. Why would these people that are developing the tests want to design a test that would hurt their own sex?

Ms. **Rosser**. Well, presumbly the people who are picking those test questions are men. I mean, you don't know who is choosing which questions to use.

Mr. **Slobodin**. How do you know that? I mean, you're speculating, aren't you?

Ms. **Rosser**. I'm speculating, and so are you.

Mr. **Slobodin**. It's not based on evidence; it's speculation.

Now, have any of your studies controlled for level of preparation—for instance, comparing girls who have taken the same math courses, the same years of math, as boys?

Ms. **Rosser**. Yes, the College Board does that. They publish voluminous data on that, and they control for that.

Mr. **Slobodin**. As a matter of fact, I have that study. It showed that the male-female gap in SAT mathematical performance shrinks considerably when differences in quantitative high school course work are taken into consideration. That point may not be that important. When you start taking preparation, that could cut considerably into that disparity.

Ms. **Rosser**. But we have already brought out the fact that males and females are taking more or less four years of math, very close, in that area. Females are still doing worse.

Mr. **Slobodin**. Well, it says here, when they control for the same level of preparation, that gap is cut considerably.

I see my time is up.

Mr. **Edwards**. We have other witnesses. But we appreciate very much your valuable contribution. So thank you very much for being here today.

Panel number two is Ms. Gretchen Rigol, Executive Director of Access Services, College Board, New York, NY; and Dr. Carol Dwyer, executive director, Test Development, School and Higher Education Programs, at the Educational Testing Service in Princeton, NJ.

Miss Rigol and Dr. Dwyer, we welcome you. Do you solemnly swear or affirm the testimony you are about to give is the truth, the whole truth, and nothing but the truth?

Ms. **Rigol**. Yes.

Dr. **Dwyer**. Yes, I do.

Mr. **Edwards**. Thank you.

Miss Rigol, I believe you are first.

STATEMENTS OF GRETCHEN W. RIGOL, EXECUTIVE DIRECTOR, ACCESS SERVICES, THE COLLEGE BOARD; AND CAROL ANNE DWYER, EXECUTIVE DIRECTOR FOR TEST DEVELOPMENT, SCHOOL AND HIGHER EDUCATION PROGRAMS, EDUCATIONAL TESTING SERVICE

Ms. **Rigol**. Thank you, Mr. Chairman.

My name is Gretchen Rigol. I am executive director for Access Services of the College Board, a position I have held for 6 years. My division is responsible for directing the Admissions Testing Program, which includes the Scholastic Aptitude Test. Prior to joining the College Board, I was Director of Admissions at Pratt Institute, and also served as an admissions officer at Goucher College and Mount Holyoke College.

Founded in 1900, the College Board is a national, nonprofit association of more than 2,500 colleges and universities, secondary schools, school systems and educational associations. A description of the full range of our services and programs is attached to my testimony.

One of the original purposes of the College Board was to provide a series of common entrance examinations that would be available to students from all parts of the country, not just those few who attended well-known preparatory schools. Those first "College Boards" represented a major step toward making higher education accessible to all students—a goal that is still of paramount importance to the College Board and its member institutions.

Today, the College Board's most widely used test is the SAT. A 3-hour, multiple choice test, the SAT measures developed verbal and mathematical reasoning abilities necessary to successfully pursue college-level work. It provides a common yardstick to help admissions officers understand an applicant's academic readiness for college-level work as they review transcripts from students who have taken different courses in the more than 25,000 secondary schools throughout this country.

Mr. Chairman, I welcome this opportunity you have provided to address the complex and complicated issues of fairness in testing and differences in scores among groups of test takers.

Average scores of various groups taking the SAT have been published for many years. Twenty years ago, the average SAT scores for women were slightly higher than the average scores for men on the verbal section of the SAT. This difference ranged from 2 to 7 points. But even then, women's average math scores were considerably lower than men's scores, between 41 to 47 points lower.

The first time women's average verbal scores fell below the average scores of men was in 1972. The differential in that year was two points, and for the next several years the difference fluctuated between three and six points. Then, in 1978, the difference increased to eight points, and in 1981, it became 12 points. Although there have been slight fluctuations during the past 6 years, the differences have remained between 10 and 13 points. I think it is important to remember that the total 61 point score differential that is so often mentioned includes 50 points on the math that has been evident for at least two decades.

I should emphasize that these scores are group averages and, as such, they do not reveal the different abilities of individuals within these groups. Distributions of scores reveal that the individuals within all groups display the full range of developed abilities, from highest to lowest.

Average score differences are of great interest, but I would like to state now that, based on the best available data, we do not believe these differences are caused by bias in the tests themselves. In many ways, this hearing and the ongoing investigations about differences in score performances are similar to the work undertaken in the seventies to help educators understand the overall score decline that began during the late sixties. Just as the Advisory

Panel on the SAT Score Decline rejected the notion of any one single cause for the overall decline in SAT scores, I suspect that there are probably numerous factors involved in this inquiry.

What are some of the possible reasons for the score differences? One is that the number of women taking the SAT increased significantly in the early seventies, just as the number of men decreased. The growth in the numbers and proportions of women SAT takers from only 44 percent in 1964 to 50 percent in 1975, and 52 percent since 1981, corresponds exactly to the time periods in which the scores of women declined. This past year, there were about 40,000 more women than men who took the SAT. When dealing with average scores on tests that are taken by self-selected populations, rather than balanced samples of students at all ability levels, it is usual for higher proportions of test takers to result in lower test scores.

We believe that the increase in the number of women taking the SAT—presumably because more women are considering a college education—should be regarded positively.

Another reason for the score differences between men and women is also related to shifts in population characteristics. The larger number of female test takers have, in recent years, included more women from racial and ethnic minorities. For example, of the nearly 80,000 black students who took the SAT in 1985, 60 percent were women. Women represented 55 percent of the American Indians who took the test, and the percentage of women in the Puerto Rican and Mexican-American groups were 54 and 53 percent, respectively. It is well recognized that the educational opportunities available to many of these minority students are not the same as those available to white students.

Parents of the females taking the SAT had slightly less formal education than the parents of the male students taking the tests, and female students tended to come from families with lower median incomes. We also know that, as a group, the women taking the SAT were less likely to have followed an academic or college preparatory program in high school and that, on average, they took fewer years of study in academic subjects.

The courses women take in high school, as we have been discussing, are also a factor in explaining some of the score differentials. For example, the more math students take in high school, indeed, the better they do on the SAT math section. The fact that women take fewer math courses than men probably explains a large part of the 50 point difference in SAT math scores. Women also take fewer courses in the physical sciences.

I am sure that you all share my concern that many young women are not encouraged to take more math and science courses and that so few consider scientifically oriented career paths. It is difficult to know exactly how much of the score difference in math is related to these unfortunate social influences, but I personally am convinced that there is no inherent difference between men and women which preclude women from excelling in the areas of mathematics and sciences.

Although the difference in average verbal scores is not as great as the difference in math, it is more difficult to suggest explanations for this 11-point difference. Some is undoubtedly related to the population differences described earlier. We have examined the test specifications and many of the items on previous editions of the test and have found no systematic explanations for the difference. There are questions on which women do less well than men, but there are also questions on which women do better. Usually these items are neutral in content and do not suggest any plausible reasons for the differential performance. There have also been numerous changes in the test content over the years; however, none of those changes coincide with the times when there were significant shifts in scores.

During the past few weeks, there have been allegations that the test is constructed to intentionally produce scores at different levels for various subgroups. I would like to state categorically that this is absolutely untrue. The College Board is committed to administering fair, effective and equitable tests. Our members would accept no less.

We use a variety of methods to detect or evaluate for the possibility of any potential bias in our tests. Among them are numerous reviews and statistical analyses that are described in my statement and that will be discussed in a moment by my colleague from ETS. I should note, however, that all of this research is not done only at ETS and that the College Board makes the data available to outside researchers for their own analysis.

Another method for determining if a test is fair examines whether it predicts equally well for different groups of students. The College Board offers a validity study service to help colleges perform studies of the predictive validity of test scores and other information used in the admission and placement of students. In over 500 colleges where females and males were studied separately, the median correlation of the SAT with college freshmen grade point average was higher for women than for men. This data is included in the *ATP Guide* that was attached to my testimony. In other words, the SAT has proved to be a more accurate predictor for women than for men.

Much has been said recently about the so-called "under-prediction" by the SAT of women's college grades. The data on which this statement is based comes from a research report published by the College Board. The data show that in the particular studies analyzed in this report, women's actual college grades were four one-hundredths of a grade point higher than their predicted grades using a combination of high school academic record and the SAT—not just the SAT alone.

More significant, however, is the fact that the prediction equations used in that study were based on the sexes combined. If a prediction equation based on women alone had been used, under- and overprediction is eliminated. Our Guidelines on the Uses of College Board Test Scores and Related Data specifically encourage colleges to consider separate predictions of college grades based on sex, race, academic program, and so forth.

There have been recent suggestions that women are being unfairly denied admission to higher education because of their SAT scores. The evidence is just the contrary. The increase in the number of women taking the SAT over the last 20 years has been mirrored in their college-going rate. More women seek entrance to and attend college than men. For example, total enrollments in higher education in 1983, the latest year for which statistics are available, were 52 percent female and 48 percent male. This is identical to the proportions of females and males that took the SAT in that year.

We have collected data for 77 different colleges that accepted fewer than 50 percent of their applicants. Mr. Chairman, you may be interested to know that Stanford was one of those colleges, where the acceptance rate was 13 percent for men, but 15 percent for women. Overall, the acceptance rate for women at these 77 colleges was 34 percent. For men, it was 33 percent. Clearly, women are finding that the doors to even the most selective colleges and universities are open to them.

In conclusion, I would like to reiterate our commitment to administering fair and equitable tests. The review process, statistical approaches, and validity studies are continually examined, refined, questioned and analyzed. With changing demographics and the diversity of test takers, questions of bias and fairness will become even more significant and more of a challenge.

It is a tragic fact of American life that educational opportunity is still not equal for all students. The educational deficit experienced by many minority and disadvantaged students will neither disappear nor be overcome simply by attributing different levels of performance on tests to bias. Although the educational opportunities available to women are comparable to those available to men at similar school settings, the fact that women are not always encouraged to take full advantage of these opportunities cannot be overlooked. Tests help reveal differences, and it is essential that we work together to try to eliminate the cause of these differences, rather than blame the "messenger" for bringing the reality of this educational deficit to our attention.

Thank you, Mr. Chairman.

[The statement of Gretchen W. Rigol, with attachments, follow:]

The College Board
1717 Massachusetts Avenue, N.W., Washington, D.C. 20036
(202) 332-7134

Washington Office

Testimony
on the Use of
Standardized Tests, and Race and
Gender Performance Differences on Such Tests

before the
Subcommittee on Civil and Constitutional Rights
Committee on the Judiciary
U.S. House of Representatives

Gretchen W. Rigol
Executive Director
Access Services
The College Board

April 23, 1987

A nonprofit educational association serving students, schools, and colleges through programs designed to expand educational opportunity

SUMMARY

It is the College Board's position that differences in average scores on the Scholastic Aptitude Test (SAT) are not caused by bias in the tests themselves. The primary purpose of the SAT is to measure the developed verbal and mathematical abilities of individuals and it does this accurately. The SAT is a carefully constructed test subjected to numerous fairness reviews and statistical analyses designed to eliminate the possibility of ethnic, racial, cultural, and gender bias.

Possible reasons why some groups -- such as women -- score lower than others include the following: First, the growth in the numbers of women test takers from only 44 percent in 1964 to 52 percent since 1981 corresponds with the period in which the scores of women declined. It is usual for higher proportions of test takers to result in lower test scores. Second, the women taking the SAT increasingly have included greater numbers who are less educationally and economically advantaged than their male counterparts. Third, on the average, women take fewer years of study in academic subjects than men; are less likely to have followed an academic or college preparatory program in high school; tend to come from families with lower median incomes; and have taken fewer mathematics and physical science courses in high school than men.

The so-called "underprediction" by the SAT of women's college grades results from using prediction equations based on the sexes combined. When a prediction equation is used based on women alone -- which is recommended by the College Board -- under-prediction is eliminated. In validity studies at over 500 colleges where females and males were studied separately, the SAT has been a more accurate predictor for women than men.

Neither is the SAT a barrier for women seeking postsecondary education. More women seek entrance to, and attend, college then men. Total enrollments in higher education in 1983 were 52 percent female and 48 percent male -- the same proportion that took the SAT that year. Last year, at 77 of the nation's most selective colleges, the acceptance rate was 34 percent for women and 33 percent for men.

It is a tragic fact of American education that educational opportunity is still not equal for all students. In addressing the issue of differences in score performance, it is important to look beyond test scores to the widely divergent educational experiences and backgrounds of the test takers. Tests such as the SAT help reveal these differences. It is essential that efforts be made to eliminate the cause of these differences, rather than blame the messenger for highlighting the reality of this educational deficit. Tests continue to remind us of an unfinished social agenda, and SAT scores reflect educational reality, rather than educational ideals.

Mr. Chairman, my name is Gretchen Wyckoff Rigol, and I am Executive Director for Access Services at the College Board, a position I have held for six years. My division is responsible for administering and directing the Admissions Testing Program, which includes the Scholastic Aptitude Test (SAT), and the Achievement Tests, as well as the Preliminary Scholastic Aptitude Test/National Merit Scholarship Qualifying Test (PSAT/NMSQT) and other related services. In my current capacity, I work with representatives from College Board member institutions and other users of our services to review policies and procedures relating to these programs. Prior to joining the College Board I was Director of Admissions at Pratt Institute and also served as an admissions officer at Mount Holyoke and Goucher Colleges.

Founded in 1900, the College Board is a national, non-profit association of more than 2,500 colleges and universities, secondary schools, school systems, education associations, and agencies. One of the purposes of the College Board is to assist students who are making the transition from high school to college through guidance and admissions programs and to provide them and the institutions to which they are applying with placement, credit by examination, and financial aid services. A description of the full range of our programs is attached to this testimony.

One of the original purposes of the College Board was to provide a series of common entrance examinations that would be available to students from all parts of the country, not just those who attended a few well-known preparatory schools. Those first "College Boards" represented a major step toward making higher education accessible to all students---a goal that continues to be of paramount importance to the College Board and its member institutions.

College admissions has changed in many ways since the beginning of this century, and College Board tests have played a role in opening college doors for large numbers of students. The Admissions Testing Program continues to enable colleges and universities to identify talented students from vastly diverse backgrounds and recruit those with academic potential. (Another example of the College Board's commitment to promoting access to higher education is the College Scholarship Service. In the 1950's the College Board membership responded to the need for a more equitable distribution of financial aid by pioneering procedures for awarding such aid according to financial need, a move that also increased educational opportunities for the less affluent and raised the level of participation in postsecondary education of minority students.)

Today, the College Board's most widely used test is the Scholastic Aptitude Test (SAT), which is taken by more than one and one-half million college-bound students every year. A three-hour, multiple-choice test, the SAT measures developed verbal and mathematical reasoning abilities necessary to pursue college-level work successfully. It provides a common yardstick to help admissions officers understand an applicant's academic readiness for college-level work as they review transcripts from students who have taken different courses in the more than 25,000 secondary schools throughout this country.

As a former admissions officer, I can assure you that it is not always easy to interpret what the actual content of a course might have been, let alone what the grading practices are in a particular school. We all know that an "A" from a certain teacher in one course might be quite different from an "A" in a different course or from a different teacher. In addition, some schools provide additional weight to certain honors or advanced level courses, while others do not. And at many colleges, nearly all of the applicants have very high grade-point averages, making it even more difficult to differentiate among applicants.

Although grade inflation appears to have slowed down during the past few years, the average high school grade-point average for the Class of 1985 was still slightly higher than a B average (3.03 on a 4.0 scale). Therefore, results from the SAT or other national standardized tests, given under similar conditions to all students, provide admissions officers with an objective context from which to view other information they have about their applicants.

It is important to remember, however, that the SAT is only one of the factors considered by colleges in making admissions decisions. Despite the limitations of information about a student's secondary school background (such as grades, courses-taken and class rank), the high school record is still given more weight than any other criteria by most colleges and universities in making admissions decisions.

But, SAT or any other test scores have limitations too. For one thing, they are not precise measures. The current score reports sent to students, as well as to the colleges they designate, show how scores should be viewed as ranges around the numerical scores that are also reported. There are also many other qualities that colleges may value and that might be important to successful performance in college that the SAT does not measure. For example, the SAT does not reflect special talents or leadership qualities nor can it predict the academic motivation or self discipline a student may bring to the collegiate environment. It cannot predict every type of performance nor measure every kind of background that may be of interest to a college. But SAT scores do provide one more piece of useful information to help both a college and a student assess how well that student might do at that particular institution, particularly when considered in the context of other relevant information about the test taker and the institutional environment.

Representatives of College Board member institutions who serve on various advisory Councils have developed a series of *Guidelines on the Uses of College Board Test Scores and Related Data* which enumerate the proper uses of tests and highlight practices deemed inappropriate. These *Guidelines*, which are included in the *ATP Guide for High Schools and Colleges*, are widely distributed to schools and colleges. Test scores, according to these *Guidelines*, should be used as "supplemental to the secondary school record and other information about applicants in assessing their ability to undertake college-level studies, recognizing that a combination of predictions is almost always better than a single prediction." To further encourage the proper use of test scores, the College Board sponsors professional training for school counselors and college admissions officers and disseminates a variety of publications and audio-visual materials.

Mr. Chairman, I welcome the opportunity you have provided to address the complex and complicated issue of fairness in testing and differences in scores among groups of test takers. As you have requested, my testimony today will focus primarily on score differences between men and women on the SAT and, secondarily, on racial and ethnic differences.

SAT scores are reported separately on a scale of 200 to 800 for both the verbal and mathematical sections of the test. To help put the discussion that follows in context, I should mention that the overall average SAT-verbal score for the Class of 1986 was 431 and the overall average SAT-math score was 475.

Average scores of various groups taking the SAT have been published for many years. Twenty years ago, the average SAT scores for women were slightly higher than average scores for men on the verbal sections of the SAT (ranging from 2 to 7 points), but even then, women's average mathematical scores were considerably lower than men's average scores (ranging from 41 to 47 points in the late 1960's). Although these differences were noted and were well-known to educators, I do not believe any definitive reasons were discovered to explain why, during that time, women performed slightly better than men on the verbal section and scored considerably lower on the mathematical section of the SAT.

Although the gap between men's and women's math scores have remained about 40 or 50 points for the past two decades, there has been a gradual change in the relative performance of men and women on the verbal section of the test during this period. The first time women's average verbal scores fell below the average verbal scores of men was in 1972. The differential in that year was 2 points and for the next several years the difference hovered between 3 and 6 points. Then in 1978, the difference became 8 points and in 1981 it became 12 points. Although there have been slight fluctuations during the past six years, the differences have remained between 10 and 13 points. The optimist (and perhaps the feminist) in me would like to suggest that the past three years that have seen the score differential move from 13 to 12 and last year to 11 points is perhaps a trend that will reverse these differences, but perhaps that's merely wishful thinking. Nonetheless, I think it is important to remember that the total 61 point score differential includes 50 points on the math section that has been evident for at least two decades.

The most comprehensive reports about SAT takers, including information by sex and by racial/ethnic group, are a series called Profiles, College Bound Seniors. The most recent report in this series describes the high-school graduates of 1985, and the data provided below are taken from that publication. The numbers in parentheses indicate the difference between the average scores for men and women of each racial/ethnic group.

Table 1. 1985 SAT Scores by Sex and Ethnic Group

	---SAT Verbal---			-SAT Mathematical-		
	Males	Females		Males	Females	
American Indian	401	384	(17)	454	406	(48)
Black	354	341	(13)	394	364	(30)
Mexican American	393	373	(20)	452	402	(50)
Asian American	406	401	(5)	540	496	(44)
Puerto Rican	385	363	(22)	435	381	(54)
White	454	444	(10)	515	468	(47)
Other	398	384	(14)	478	419	(59)
Total Respondents	437	425	(12)	499	452	(47)

When the average scores of males and females from the different racial/ethnic groups are reviewed, it is clear that male/female differences are not constant across all groups. On the verbal sections, Asian American men and women show the smallest differences, while thelargest differences are evident between men and women with Hispanic backgrounds. When looking at average SAT-math scores, Black women have the smallest difference when compared with Black men, with larger score differences apparent for all other groups. These data illustrate the complexity of the issue.

I should emphasize that these scores are group averages, and as such they do not reveal the different abilities of individuals within those groups. Distributions of scores reveal that the individuals within groups (whether that group be based on sex or racial/ethnic group) display the full range of developed abilities, from highest to lowest.

The SAT is not the only test that shows score differences among the different groups taking the test, particularly differences between male and female scores. The same trend exists for American College Testing (ACT) Program scores. The ACT includes separate scores in four areas: English Usage, Mathematics Usage, Social Science Reading, and Natural Science Reading and is scored on a scale from 1 to 36. Between 1970 and 1984, the advantage of women on the English score declined from an average of 1.8 ACT score points to 1.1. Similarly, the advantage of men on the other three ACT tests and on the ACT composite grew over the same period of time. For example, from 1970 to 1984 on the Social Studies Reading subscores, the advantage of males climbed from 1.3 to 1.6 and on the Natural Science Reading from 1.6 to 2.5. Data from the National

Assessment of Educational Progress (NAEP) and other standardized tests show similar trends, suggesting a deterioration of women's average scores in relation to average scores of men.

Average score differences are of great interest, but I would like to state now that, based on the best available data, we do not believe these differences are caused by bias in the tests themselves. We are pleased that this Subcommittee has provided an open forum to discuss and examine the issues. We invite the members of the Subcommittee to ponder with us and other educators the dilemma of trying to explain differential performance and changes over time. We would be less than honest if we claimed to have all the answers. We can offer some hypotheses, but we continue to question research and our conclusions. In many ways, this hearing and the ongoing investigations about differences in score performances are similar to the work undertaken in the mid-1970s to help educators understand the overall score decline that began during the late 1960s. Just as, in 1977, the Advisory Panel on the SAT Score Decline, headed by former Secretary of Labor Willard Wirtz, rejected the notion of any one single cause for the decline in SAT scores, I suspect that there are probably numerous factors involved in this inquiry.

Why, then, do some groups of students score lower than others? In addressing this issue, we must look beyond the test scores to the educational experiences and backgrounds of the test takers. It is a tragic fact of American education that educational opportunity is still not equal for all students. The educational deficit experienced by many minority and disadvantaged students will neither disappear nor be overcome simply by attributing different levels of performance on tests to bias.

Although the educational opportunities available to women are comparable to that available to men in similar school settings, the fact that women are not always encouraged to take full advantage of these opportunities cannot be overlooked. Tests help reveal differences and it is essential that we work together to try to eliminate the cause of these differences, rather than blame the messenger for bringing the reality of this educational deficit to our attention.

What are some of the possible reasons for these score differences?

One is that the number of women taking the SAT increased significantly in the early 1970s, just as the number of men decreased. The growth in the numbers and proportions of women SAT takers from only 44% in 1964 to 50% in 1975 and 52% since 1981 corresponds exactly to the time periods in which the scores of women declined. This past year, there were about 40,000 more women than men who took the SAT. When dealing with average scores on tests that are taken by self-selected populations, rather than balanced samples of students at all ability levels, it is usual for higher proportions of test takers to result in lower test scores. For example, if only the top 10% of a group of students takes a test, the average scores for this group would probably be higher than a much larger group of students who represent a wider range of abilities.

We believe that the increase in the number of women taking the SAT -- presumably because more women are considering a college education -- should be regarded positively. It is indicative of changing mores and social patterns that have heightened women's expectations about their educational options and their careers.

Another reason for SAT score differences between men and women is also related to shifts in population characteristics of students taking the test. As the table below indicates, the percentage of women from the various racial/ethnic groups is not equal. This data is also taken from Profiles, College-Bound Seniors, 1985.

Table 2. 1985 College-Bound Seniors: Number of Students by Ethnic Group and Sex

	Total Number	Percent of Total	Total Number Female	Percent Female
American Indian	4,642	0.5	2,563	55.2
Black	79,556	8.9	47,866	60.2
Mexican American	19,526	2.2	10,395	53.2
Asian American	42,637	4.8	20,959	49.2
Puerto Rican	11,077	1.2	6,000	54.2
White	715,773	80.0	373,694	52.2
Other	21,555	2.4	10,839	50.3
Total Respondents	894,766	100.0	472,316	52.8

The larger numbers of female test takers have, in recent years, included more women from racial and ethnic minorities. For example, of the nearly 80,000 Black students who took the SAT in 1985, 60% were women. Women represent 55% of the American Indians who took the test and the percentage of women in the Puerto Rican and Mexican American groups were 54% and 53% respectively. As I have noted earlier it is well recognized that the educational opportunities available to many of these minority students are not the same as those available to white students. Indeed, within these minority groups, females are often at a further disadvantage.

We also know that, as a group, the women taking the SAT were less likely to have followed an academic or college preparatory program in high school than men and that, on the average, they took fewer years of study in academic subjects than males. Parents of the females taking the SAT had slightly less formal education than males and the females tended to come from families with lower median incomes. Although females do come from families of all educational and economic backgrounds and many have taken rigorous academic programs, as a group they are not quite as well prepared nor are they from homes as advantaged as the smaller number of male test takers. Although I am gratified that more women from backgrounds that traditionally have not considered college are pursuing higher education, these data also raise concerns that such low proportions of minority males are considering college.

The courses women take in high school also are a factor in explaining some of the score differentials of male and female students. For example, the more mathematics women study in high school, the better they do on the SAT math section. The fact that women take fewer mathematics courses, on average, than men probably explains a large part of the difference in the SAT-math scores. Women also take fewer courses in the physical sciences.

I am sure that you share my concern that many young women are not encouraged to take more math and science courses and that so few consider scientifically-oriented career paths. It is difficult to know exactly how much of the score difference in math is related to these unfortunate social influences, but I personally am convinced there is no inherent difference between men and women which preclude them from excelling in the areas of mathematics and sciences. Tests continue to remind us of an unfinished social agenda, and SAT scores reflect educational reality, rather than educational ideals.

Although the difference in average verbal scores is not as great as the difference in math, it is more difficult to suggest explanations for the 11-point difference in SAT verbal scores between men and women. Some of the difference is probably related to population differences described earlier. We have examined the test specifications and many of the items on previous aditions of the test and have found no systematic explanations for the difference. There are questions on which women do less well than men and there are also those on which women do better. Usually such items areneutral in content and do not suggest any plausible reasons for the differential performance. There have been numerous changes in test content over the years; however, none of these changes coincide with the times when there were significant shifts in scores.

During the past few weeks there have been allegations that the test is constructed to intentionally produce scores at different levels for various subgroups. I would like to state categorically that this is absolutely untrue. The College Board is committed to administering fair, effective and equitable tests. Our members would accept no less. As a result of substantial research efforts, the Board believes that the SAT reflects accurately the developed verbal and mathematical abilities of the individuals who take it, regardless of their sex or racial or ethnic background.

There are three basic methods used by the College Board to detect any potential biases: reviews by numerous committees and panels, statistical analysis and validity studies. It is significant to note that these efforts date back to the 1920's -- the very early days of the SAT. "Precautionary studies" of the test performance of males and females, for example, were conducted from the beginning, reflecting the Board's strong concern in this area.

Since the late 1940s, the SAT and most other College Board tests have been developed by the Educational Testing Service (ETS). ETS shares the College Board's commitment to offer tests that are not influenced by extraneous cultural, ethnic or social factors and toward this end employs numerous procedures to ensure the tests are free from any such influences.

Current practices require that each new College Board test undergo a sensitivity review to identify and eliminate ambiguity or potentiallyoffensive material based on race, sex, and cultural background. Sensitivity reviewers are trained to ensure thorough knowledge of the review process and consistent application of review criteria. They are selected on the basis of their ability to perceive potentially offensive material; to review tests from multiple perspectives, not simply from the viewpoint of one group or social/political philosophy; and to cover key subject areas such as humanities and social sciences. During the past year, sensitivity reviewers of new editions of the SAT included 14 women, four Blacks, two Asian Americans, and two Hispanics.

In addition to these formal sensitivity reviews, each College Board test is thoroughly reviewed by the high school and college faculty serving on the SAT Committee, as well as external review panels for both the verbal and mathematical sections. All committees and special review panels are selected from among a cross section of backgrounds, including minority representation and women, academic disciplines, institutional affiliations and geographical representation.

Since the late 1970s, each new form of the SAT also includes at least one passage dealing with minority issues. New tests are also reviewed carefully to ensure that an appropriate variety of references to women and minorities are included throughout the test. These content specifications are most apparent in the reading comprehension passages and in the sentence completion items, which have more text than the analogy, antonym and mathematics questions.

Finally, it should be noted that women have been involved in all aspects of the development of the SAT for several decades. Since 1973, a woman has been the primary test development specialist for the verbal test. Of the ETS staff members who spend significant amounts of timeworking on the SAT, there are 11 women and 7 men working on the verbal sections and 6 women and 3 men working on the mathematics sections. There currently are 15 women and 10 men who serve as outside item writers for SAT-Verbal; 9 women and 9 men outside of ETS who write mathematics questions.

Statistical methods that consider the performance of groups on individual questions or clusters of questions are also used to ensure test fairness and to detect any possible bias. We make available the raw data to the research community in an ongoing effort to ascertain the specific causes of differential performance on tests. For example, a Public Use Sample data tape, containing all of the information about test candidates from the largest administration of the SAT each year, is offered as a regular service. The College Board welcomes external research on the issue and invites the public to analyze or reanalyze the data.

The College Board has conducted numerous studies to examine how different groups perform on various SAT items. During the past few years, item fairness studies have been conducted on the basis of sex, ethnicity, educational background of parents, and level of English proficiency. An article published by the College Board in 1981, "The SAT in a Diverse Society: Fairness and Sensitivity," summarizes the kinds of analyses that result from these studies. The purpose of these statistical studies is to monitor differential performance in order to (1) ensure that the SAT remains appropriate over time for major subgroups of the candidate population, and (2) identify possible content factors related to differential performance. If the analysis identifies any questions with large differentials, further analysis to identify the causes is conducted.

ETS has recently developed a new statistical procedure that holds promise in further detecting any potential bias in our tests. Known as "differential item functioning" (DIF), this statistical procedure matches people of the same ability level before comparing their performance on test questions. The assumption is that individuals of similar knowledge and skill should have similar chances of answering a question correctly without regard for their race, sex or ethnic background. The statistics, thus, compare the performances of majority and minority students, and men and women of similar ability. Research on DIF is continuing and plans are being developed for its use at various stages in the test development process.

The third method for determining if a test is fair examines whether it predicts equally well for different groups of students. Predictive validity is the measure of a test's effectiveness in predicting the academic performance of a student in college. The College Board offers a Validity Study Service, without cost, to colleges that wish to evaluate how well their admissions data predict the academic performance of their enrolled students. The service provides assistance in performing studies of the predictive validity of high school records, test scores and other information used in the admission and placement of students.

Over 1,300 of these validity studies have been conducted by colleges and universities in the past several years to determine whether the test scores predict the expected outcome in the freshmen year. These studies also help indicate the relative weight that should be given to SAT scores and other data (such as high school grade-point average or high school rank) in the admissions process.

These validity studies indicate that in over 500 colleges where females and males were studied separately, the median correlation of the SAT with college freshman grade point average was higher for women than for men as indicated in Table 16 of the AIP Guide. Thus, the SAT has proved to be a more accurate predictor for women than men.

Much has been said recently about the so-called "underprediction" by the SAT of women's college grades. The data on which this statement is based comes from a research report authored by Mary Jo Clark and Jerilee Grandy and published by the College Board. These data show that in the particular studies analyzed, women's actual college grades were four one-hundredths of a grade point higher than their predicted grades using a combination of high school academic record and the SAT, not the SAT alone. More significant is the fact that the prediction equations used in that study were based on the sexes combined. However, as noted above, if a prediction equation based on women alone was used, under- and over-prediction would be eliminated. Our Guidelines on the Uses of College Board Test Scores and Related Data specifically encourage colleges to consider separate predictions of college grades based on gender, race, and ethnicity.

Data about the validity of the SAT for various racial groups have also been studied by both the College Board and others. After two years of intensive examination by experts, the 1982 National Academy of Sciences study, Admissions Testing in Higher Education, concluded "that predictions made from test scores are as accurate for black applicants as for majority applicants; there is only scanty evidence available for other minority groups. Subgroup differences in average ability test scores appear to mirror like differences in academic performance as measured by course grades. In this sense, the tests are not biased."

Before concluding, I would like to address briefly an issue that has received much attention recently: the Empire State Scholarship (New York State) awards and the disparate number of male recipients. You may be aware that these scholarships are awarded on the basis of SAT or ACT scores. The College Board does not support the use of SAT scores as the sole criterion in any decision making process -- even in admissions for which the test is designed. But the issue here is one of social values and whether the intent of the scholarship program is to recognize scholastic ability (or any other factor) regardless of the composition of population competing for such awards or whether such awards should be apportioned among various subgroups. The designation of awards by subpopulations is frequently a major part of many scholarship programs. This might be allocating a certain number of awards for congressional districts, a certain percentage for men and women or for various racial/ethnic groups. Clearly, no single test will automatically result in such desired allocations.

There have also been recent suggestions that women are being unfairly denied admission to higher education because of their SAT scores. The evidence is just the contrary. The increase in the number of women taking the SAT over the last twenty years has been mirrored in their college going rate. More women seek entrance to, and attend, college than men. For example, total enrollments in higher education in 1983, the latest year for which statistics are available, were 52% female and 48% male. This is identical to the proportions of females and males that took the SAT in that year. Data provided by colleges for the Annual Survey of Colleges, which forms the basis for the *College Handbook*, includes acceptance rate information separately for men and women for 77 different colleges that accepted fewer than 50 percent of their applicants to their fall 1985 class. Overall, the acceptance rate for women was 34%, for men it was 33%. Clearly, women are finding that the doors to colleges and universities are open to them.

In conclusion I would like to reiterate our commitment to administering fair and equitable tests. The review process, statistical approaches, and validity studies I described earlier are continually examined, refined, questioned and analyzed. With changing demographics and the diversity of test takers, questions of bias and fairness will become even more significant and more of a challenge. This week the American Educational Research Association is holding its annual meeting here in Washington. A look at its program is illustrative of the importance researchers and test developers place on questions of fairness in testing and the analytical methods to achieve that goal.

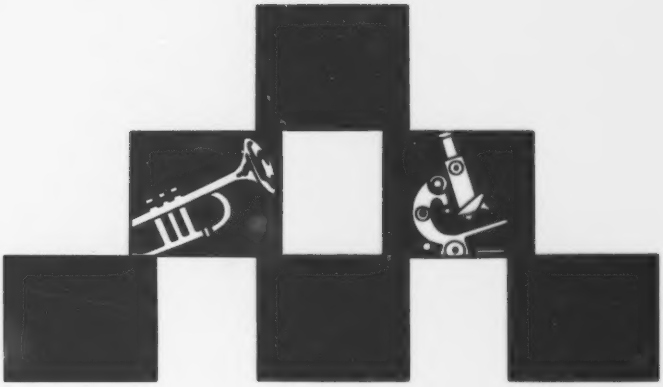
Thank you, Mr. Chairman. This concludes my prepared statement. I will be happy to answer any questions you might have.

1986-87

ATP GUIDE

For High Schools and Colleges

SAT® and
ACHIEVEMENT TESTS



THE COLLEGE BOARD

Contents

[illegible]

To Ask a Question about Services for Students

WRITE: College Board ATP, CN 6200, Princeton, NJ 08541-6200

PHONE: Monday to Friday
Princeton, NJ **609-771-7600** 8:30 a.m. to 9:30 p.m. Eastern time
Berkeley, CA **415-849-0950** 8:15 a.m. to 4:30 p.m. Pacific time

---

Contact Your College Board Regional Office (back cover) for Information about:

- using SAT, TSWE, and Achievement Test scores
- workshops and other instructional programs for staff development
- score report formats (magnetic tape, labels, etc.)
- information about fee waivers
- conducting or updating validity studies
- testing on campus

Publications cited in this *Guide* can be ordered from:

College Board ATP
CN 6212
Princeton, NJ 08541-6212

For a more technical discussion of ATP tests than is given in this *Guide*, see *The College Board Technical Handbook for the Scholastic Aptitude Test and Achievement Tests* and the two editions of *An SAT: Test and Technical Data for the Scholastic Aptitude Tests* administered in March 1980 and April 1981.

This guide to the ATP and related services is prepared for high schools and colleges by Educational Testing Service, which develops and administers the tests of the ATP for the College Board. Up to five additional copies will be supplied to high schools and colleges free upon request. Copies in quantities of more than five can be ordered at $1.00 per copy.

The Admissions Testing Program is a program of the College Board, a nonprofit membership organization that provides tests and other educational services for students, schools, and colleges. The membership is composed of more than 2,500 colleges, schools, school systems, and education associations. Representatives of the members serve on the Board of Trustees and advisory councils and committees that consider the programs of the College Board and participate in the determination of its policies and activities.

The College Board and ETS are dedicated to the principle of equal opportunity, and their programs, services, and employment policies are guided by that principle.

Copyright © [illegible] by College Entrance Examination Board. All rights reserved.

College Board, Scholastic Aptitude Test, SAT, [illegible] are registered trademarks of the College Entrance Examination Board.

Library of Congress Catalog Card Number [illegible]

BEST COPY AVAILABLE

The Admissions Testing Program

The College Board Admissions Testing Program (ATP) is designed to assist students, high schools, colleges, universities, and scholarship programs with postsecondary educational planning and decision making and to provide a channel of communication between students and these institutions. Because the subject matter of high school courses as well as grading standards vary widely, the ATP tests have been developed to provide a common standard against which students can be compared.

The ATP consists of the Scholastic Aptitude Test (SAT), the Test of Standard Written English (TSWE), the Achievement Tests, and the Student Descriptive Questionnaire (SDQ). Closely related to the ATP are the Student Search Service, the Summary Reporting Service, and the Validity Study Service.

The tests are developed with the assistance of experienced high school and college teachers who set specifications and review the content. Meticulous care goes into the writing, pretesting, research, and evaluation of each of the tests. Rigorous adherence to standards is maintained throughout the administering, scoring, and reporting phases of the program. There are no age or grade restrictions for taking the tests, and all of them are available to students with handicaps. Most students take the tests during national administrations in their junior and senior years of secondary school. Some colleges have special arrangements for testing students who have not tested before applying. (See "Testing on Campus," page 8.)

The Scholastic Aptitude Test

The SAT is a 2½-hour multiple-choice test that measures developed verbal and mathematical reasoning abilities related to successful performance in college. It is intended to supplement the secondary school record and other information about the student in assessing readiness for college-level work.

The Test of Standard Written English

The TSWE is a 30-minute multiple-choice test administered with the SAT. The TSWE measures students' ability to recognize and use standard written English. Scores can be used by colleges to help place students in appropriate freshman English courses.

The Achievement Tests

Achievement Tests are designed to measure knowledge and the ability to apply that knowledge, in specific subject areas. Achievement Tests are independent of particular textbooks or methods of instruction. Although types of questions change little from year to year, the tests do evolve to reflect general trends in high school curriculums.

Those colleges requiring Achievement Tests use them in selecting students for admission, for course placement, or both. Some colleges specify the tests to be taken or the subject areas; others allow applicants to choose.

Achievement Tests are given in English Composition, Literature, American History and Social Studies, European History and World Cultures, Mathematics Level I, Mathematics Level II, French, German, Hebrew, Latin, Spanish, Biology, Chemistry, and Physics. All are one-hour multiple-choice tests with the exception of the December version of the English Composition Test, which is composed of 40 minutes of multiple-choice questions and one 20-minute essay assignment.

The Student Descriptive Questionnaire

The SDQ, which is answered by about 90 percent of the students when they register for the SAT or Achievement Tests, contains questions about the students' background, high school courses and other educational and extracurricular experiences, and plans for college study.

The SDQ contributes to guidance and admissions by enabling individual students to present a broader picture of themselves to colleges than is conveyed by test scores alone. Students' answers to the SDQ and their scores on the tests are the primary sources of the tables in the ATP Summary Reports. The SDQ is also one of the sources of information used in the Student Search Service to identify students with specific characteristics designated by participating colleges and scholarship agencies.

Note: Because the SDQ was new in 1985-86, students who completed the SDQ prior to October 1985 and who wish to have SDQ information reported to colleges must complete a new SDQ if they register to test again or submit an SDQ Update Form, included in summer bulk shipments to secondary schools.

Score Reports to Students

Approximately five weeks after the test date, students will receive their student report, called the College Planning Report, a two-page document that integrates their test scores with key information about themselves and the colleges they are considering. Designed to help students understand that their scores are only part of the college entrance picture, College Planning Reports contain students' current test scores (expressed both as numbers and as ranges), previous SAT, TSWE, and Achievement Test scores on record, national and state percentile ranks, secondary school courses and grades, and information about the colleges the students designated to receive score reports. (See pages 8 to 11 for a complete description of the College Planning Report and a sample report.)

Score Reports to High Schools

About five weeks after the test date, high schools receive, at no cost, a College Counseling Report for each student who gave that high school code number when registering for ATP tests. In addition to the students' test scores, score ranges, and percentiles, the reports contain a review of their high school programs, grades, and class rank, and college plans as indicated on the SDQ. The colleges designated by the student to receive score reports are also listed. (The College Counseling Report is described in detail on page 12 and a sample report appears on page 13.)

In addition, high schools receive:

- Two pressure-sensitive labels for each student containing current test score data useful for permanent school record retention. Scores are reported in three-digit form on these labels. SAT-verbal subscores are reported as two digits. Nonstandard administrations are also noted.
- Score rosters: alphabetical lists of students by grade level, containing current scores, sex, and ethnic group that are sent to schools after each administration. Students who were absent or whose scores are delayed for any reason will be identified only on the score roster. Cumulative rosters will be sent to schools in January and June.

High schools and school districts also may purchase magnetic tapes containing each student's SAT and Achievement Test scores. These reports are sent after every test administration and include scores from previous test dates as well as SDQ responses.

Score Reports to Colleges

About four weeks after each test date, colleges receive reports for all students who indicated that their test scores be sent to it. Colleges may request reports in one of the following formats without charge. Additional formats are available for a fee.

- The College Admissions and Advising Report: contains, in addition to the student's current and previous test scores and academic profile, a student information section useful for placement and advising purposes. (See page 12 and pages 14 to 17 for a complete description and sample report.)
- Magnetic tape: contains all the information on the paper reports.
- Pressure-sensitive labels: give current and previous test scores, reported as two digits.
- Pressure-sensitive mailing labels: contain only the name and address of the student. (Available only as a second option.)

Further information about score report formats is included in the booklet *1986-87 ATP Score Report Options*, which colleges receive during the summer.

After each test administration, colleges that received scores for 25 or more students will receive summary statistics on scores reported. These summary statistics provide distributions, means, and standard deviations for the students who requested that their scores be sent. In addition, colleges also will receive a Trend Data Report that includes comparative information on the number of score reports sent to that institution during the current and previous year.

The Student Search Service

The Student Search Service assists colleges and governmental scholarship programs in identifying students with certain characteristics, based on information the students provide in the SDQ. Only postsecondary institutions, groups, associations, and consortia eligible to be included by the U.S. Department of Education in its current *Education Directory, Colleges and Universities*, government scholarship agencies, and postsecondary institutions, groups, associations, and consortia that are members of the College Board are eligible to use the service. Institutions participating in the Student Search Service specify student characteristics in which they are interested, such as grade average, a range of test scores, intended college major, ethnic background, religious preference, and geographic location. They can then request the names and addresses of students matching these specifications. In addition, the following student information is reported to them: sex, social security number, birth date, and secondary school. Individual test scores are not reported by the Student Search Service. Participating institutions then send these students information on their programs, admissions policies, financial aid opportunities, and the like.

Students indicate their interest in being included in the Student Search Service on the Registration Form. About 80 percent do so. The Service searches its files six times a year – after the December test for students who took the SAT, three times in the spring for students who took the Preliminary Scholastic Aptitude Test/National Merit Scholarship Qualifying Test (PSAT/NMSQT) the previous fall, in the summer for juniors who took the SAT in the previous testing year, and also in the summer for juniors who took Advanced Placement Examinations. (See *Recruiting with the Student Search Service*.)

ATP Summary Reports

Each summer the College Board produces a series of reports summarizing ATP test scores and data from the SDQ for the previous year's senior class who took the tests. Summary reports are produced for secondary schools on their college-bound seniors, and for colleges on the seniors who sent ATP score reports to them. Separate reports are compiled on college-bound seniors by state, by region, and for the nation. See the *School Guide to the ATP Summary Reports* and the *College Guide to the ATP Summary Reports* for detailed descriptions.

Because a new SDQ was introduced in 1985-86, some students in the class of 1986 will have completed the old SDQ and some the new SDQ. Consequently, it will not be possible to produce a full summary report, tapes, or college reports on applicants, accepted applicants, enrolling freshmen, or persisters for the class of 1986; only abbreviated reports will be produced for this class. Beginning with the class of 1987, a new series of summary reports will be available.

The Validity Study Service

The College Board offers this service without cost to colleges that wish to evaluate how well their admissions data predict their enrolled students' academic performance. The service provides assistance in setting up studies of the predictive validity of high school records, ATP scores, and other information used in the admission and placement of students. It also determines the best weighted combination of high school grades and test scores for estimating students' freshman-year academic performance at individual institutions. (See "Using Predictive Validity Studies," page 26, and *Guide to the College Board Validity Study Service*.)

Administering the Program

Complete information regarding test dates, registration procedures, fees, special arrangements, and scoring services appears in the *Registration Bulletin*. Information to help counselors answer the questions they are asked most frequently is repeated below, with additional instructions for special situations.

Publications Sent to Schools and Colleges

During the summer, high schools receive a supply of the publications and forms listed below. Reference copies of the publications are also sent to colleges, which may order additional copies.

- *Registration Bulletin* for the SAT and Achievement Tests (for distribution to students). Contains the Registration Form, SDQ, a list of test center codes in the region, college and scholarship codes, state and county codes, and information on ATP procedures and services. There is a *New York State Edition* and an *International Edition* of the *Bulletin* in addition to the four regional editions (*Midwestern, Northeastern, Southern, and Western*).
- *Complete List of Test Centers for the SAT and Achievement Tests* (for reference). Contains all of the test center codes in all editions of the *Bulletin*.
- *Codes for SAT and Achievement Test Score Recipients* (for reference). Contains a complete list of codes for colleges and scholarship programs, Upward Bound programs, and members of the U.S. Senate and House of Representatives.
- Additional Registration Forms and envelopes (for students who register for additional test dates).
- *Taking the SAT* (for distribution to students who intend to register for the SAT). Contains examples of each type of test question with directions, explanations, and other general test-taking advice and includes a sample SAT and TSWE answer sheet, correct answers, and scoring instructions.
- *Taking the Achievement Tests* (for distribution to students who intend to register for the Achievement Tests). Explains the purpose of the tests and contains examples of each type of test question, with directions, explanations, and sample questions.
- *Using Your College Planning Report* (for staff use). Mailed to students with their score reports. Explains how students can use the information on their reports to help review their college selections. Also describes the information reported to colleges and how it is used.
- *ATP Guide for High Schools and Colleges* (for staff reference).
- Additional Report Request Forms (for students who need more than the one form they receive with the Admission Ticket).
- SDQ Update Forms (for students who wish to revise or update information they provided on the SDQ).
- School Code Poster (for display). Contains the school code number, test dates, and registration deadlines. There is a special edition of the poster for New York State and a special edition for countries other than the United States.
- Publications Shipment Notice/Reorder Form (for ordering additional copies of program publications).

Registering

Instructions for completing the Registration Form appear in the *Bulletin* and on the form itself. The registration process in the national testing program is the same for the SAT and the Achievement Tests, but students must submit a separate Registration Form for each test date. The identification information provided by the students is used to accumulate scores on score reports. Remind the students to supply identification information exactly the same way in all contacts with the ATP to avoid delay or error. Consistent identification information also helps colleges combine ATP data with other information they receive about an applicant.

Special Arrangements for Students with Handicaps

Special editions of the SAT (in regular type, large type, braille, and cassette versions) with extended testing time are available for students with documented visual, hearing, physical, or learning disabilities. Achievement Tests are available only in regular type but may be taken with extended testing time. Eligible students may take these tests at times arranged by the student and counselor.

A second option is available for taking the SAT if students have documented learning disabilities that allow use of a regular edition and a machine-scannable answer sheet but require additional testing time. In such cases, students may take the SAT at the regular national administrations in November and May, at which time they will be allowed up to one and a half hours of extended testing time. Students who test on these dates will be able to order the SAT Question-and-Answer Service (see "Verifying SAT Scores," page 7).

For any of the above arrangements, students should follow the registration procedures described in *Information for Students with Special Needs*, which can be requested from ATP Services for Handicapped Students, CN 6226, Princeton, NJ 08541-6226, or by calling 609-771-7600.

If a disability does not require special arrangements or extended testing time, students should register for the regular national program. Students with temporary disabilities (a broken arm, for example) should register for a later date in the national program unless they need to meet an application deadline.

Other Special Testing Arrangements

Special arrangements are made for students who for religious reasons cannot take the tests on Saturday or who live more than 75 miles from a regular testing center, and for service personnel who will be aboard a ship at sea on a regular test date. To request special arrangements, students must:

1. Complete and submit a Registration Form by the regular registration deadline for domestic students, or the special requests deadline for students testing outside the United States or Puerto Rico.
2. Record test center number 01-000 as the first choice in item 10 on the Registration Form.
3. Enclose with the Registration Form and fees a letter explaining the reason for the request. A statement signed by a clergy member must accompany a request for Sunday testing. A statement signed by the commanding officer must accompany requests by service personnel.

The College Board also provides special testing arrangements when school-sponsored activities (for example, an athletic competition, debate tournament, band contest) may prevent previously registered candidates from taking the tests at the regularly scheduled time or place. Students will be charged the test center change fee for switching to an alternate test date. If you know of situations that require special attention, contact College Board ATP (see inside front cover) no later than 10 calendar days before the test administration.

Fees, Fee Refunds, and Fee Waivers

Fees and fee refunds for ATP tests are listed in the *Registration Bulletin*. If students are absent from a test for which they registered, the College Board will refund the test fee (minus the service fee) if students ask for it. Fees cannot be transferred. Refund requests must be sent within two months of the scheduled test date. Service fees are not refundable.

Fee waivers are available to eligible high school juniors and seniors who need to take the SAT or Achievement Tests but cannot afford the test fee. (Fee waivers are not available to seventh, eighth, ninth, or tenth graders.) Instead of the test fee, a fee-waiver card must be submitted with the Registration Form. At the beginning of the school year, schools and special programs such as Upward Bound, and community counseling agencies are sent guidelines on eligibility and are allocated fee waiver cards on the basis of the number they used the previous year.

For information about fee waivers or additional fee-waiver cards, write or phone your College Board Regional Office. Make requests as early as possible before a test administration so that students can submit their Registration Forms with fee-waiver cards before the late registration deadline. Seniors who have never taken the tests have priority for fee waivers. Eligible students may receive only one fee waiver for the SAT and/or one for Achievement Tests, which may be used during either the junior or senior year. Fee waivers cover only the basic test fees, the SAT Question and Answer Service, and the SAT Score Verification Service; they cannot be used to cover a late fee, standby fee, additional reports, other service fees, or purchase of *The College Handbook* or *Index of Majors*.

Fee waivers are available to nationals of countries other than the United States only if they test in the U.S., Puerto Rico, or U.S. territories.

Cumulative Reporting

If students provide the same identifying information each time they register, their reports will contain current test scores and all previous SAT, TSWE, and Achievement Test scores from up to 11 previous test dates. It is not possible to send only the latest or highest test scores or separate reports for the SAT, TSWE, or Achievement Tests. If previous scores do not appear on students' reports, they should write to College Board ATP (Attention: Unreported Scores).

Additional Reports

Students may request additional reports at any time by completing an Additional Report Request Form. The fee for each additional report is $5.00. A form is enclosed with the Admission Ticket and a supply is included in summer shipments to secondary schools. Colleges or scholarship programs may order forms preprinted with their code numbers to send to applicants who have not yet submitted official score reports.

Because the SDQ was new in 1985, students tested prior to October 1985 who submit an Additional Report Request Form (and who do not plan to test again) must also complete an SDQ Update Form in order for SDQ information to be reported to colleges. A supply of SDQ Update Forms is included in summer shipments to secondary schools.

Additional SDQ Update Forms can be requested by writing or calling College Board ATP (See inside front cover).

Telephone Rush Request Service

If a student wants colleges to receive scores sooner than usual and if the scores have been processed (usually about three weeks after the test date), the student can call College Board ATP (609-771-7600) and request the rush score reporting service. Scores will be sent to the colleges and scholarship programs specified within two working days after the call. The student will receive a confirmation copy of the interim report (which contains ID information and scores only) and will be billed $15.00 for this service plus $5.00 for each report. Complete reports will be sent to the student and colleges during the next scheduled processing.

When students call, they should provide identification information as recorded on their Registration Form, the most recent test date, and the names and code numbers of the colleges and scholarship programs that should receive interim reports.

Automatic Reports to Scholarship Programs

Only students can request that their scores be sent to high schools, colleges, and scholarship programs. Scores for all seniors who attend high school in or who reside in Florida are routinely sent to their state's scholarship program. Scores for all juniors in Pennsylvania and for all juniors in Illinois who test between January 1, 1987, and June 30, 1987, are routinely sent to those states' scholarship programs. In Rhode Island, scores are sent for all seniors who take the test in November and December 1986, and January 1987. In Maryland, the most recent SAT scores are sent for all state scholarship applicants. If students who live in or attend school in one of those states do not want their scores sent to the state scholarship agency, they should notify College Board ATP, CN 6200, Princeton, NJ 08541-6200 by the appropriate date: Florida and Rhode Island, January 31, 1987; Maryland, February 15, 1987; Pennsylvania, May 31, 1987; Illinois, August 1, 1987. Students and counselors in New York State should refer to the *New York State Edition* of the *Bulletin* for a notice of the special reporting procedures used for the New York State Regents Scholarship Program.

Changing SDQ Information

Students need to complete the SDQ only once. (See note below.) If they register for a subsequent test date, they can update answers. However, they must answer the entire question because their new answer will completely replace their previous answer. For example, if they have taken a calculus course since the last time they answered the SDQ and want to update their SDQ by including this information, they must record all their previous math courses as well as calculus, even though they recorded these courses the first time they answered the SDQ. Their previous answers to all other questions will continue to be reported as they were to high schools and colleges.

Students can make changes in their SDQ at any time by calling College Board ATP (609-771-7600).

Note: Because the SDQ was new in 1985-86, students tested prior to October 1985 who wish to have SDQ information reported to colleges must complete the current SDQ (if they register to test again) or submit an SDQ Update Form, included in summer bulk shipments to secondary schools.

Verifying SAT Scores

The College Board offers two services that enable students to verify their SAT scores: The SAT Question and Answer Service and the SAT Score Verification Service. Either may be ordered up to five months after the test date.

Students who take the SAT on one of the five dates specified in the *Registration Bulletin* may order the SAT Question and Answer Service. They will receive their SAT questions, the correct answers, instructions for scoring, and a copy of their answer sheet. For other SAT dates, students can verify their SAT scores by ordering the SAT Score Verification Service. The SAT Score Verification Service includes all materials provided for the SAT Question and Answer Service except the test questions. An order form is in *Using Your College Planning Report*.

For both the Question and Answer Service and the Score Verification Service, if the score results calculated by the students disagree with the SAT scores on their score report, the students may request rescoring of their answer sheet. If rescoring confirms that an error had been made (resulting in either higher or lower scores than those originally reported), corrected reports will be sent, without charge, to all recipients of their original scores.

Preparing for the Tests

For students to perform to the best of their ability, they should know what the test is about and how it is structured, how to make the most efficient use of time limits, how to "attack" the different kinds of questions, and when an "educated" guess using partial knowledge is sensible. For this reason students should be encouraged to study the material in *Taking the SAT* and *Taking the Achievement Tests* and to complete the sample questions that are included. Schools may choose to assist students in the process through group meetings and discussion sessions to emphasize the importance of this preparation.

Special Preparation for the SAT

For more than 25 years, the College Board has sponsored research on the effects of special preparation programs on SAT score results and has supported independent investigation of this topic by others. On the basis of present knowledge, the College Board has prepared a statement to assist students in making decisions about special preparation for the SAT. A reprint of the statement follows.

- The SAT measures developed verbal and mathematical reasoning abilities that are involved in successful academic work in college; it is not a test of some inborn and unchanging capacity.
- Scores on the SAT can change as you develop verbal and mathematical abilities both in and out of school.
- Your abilities are related to the time and effort spent: short-term drill and cramming are likely to have little effect; longer-term preparation that develops skills and abilities can have greater effect. One kind of longer-term preparation is the study of challenging academic courses.

- While drill and practice on sample test questions generally result in little effect on test scores, preparation of this kind can familiarize you with different types of questions and may help to reduce your anxiety about what to expect.
- Whether longer preparation, apart from that available to you within your regular high school courses, is worth the time, effort, and money is a decision you and your parents must make for yourselves; results seem to vary considerably from program to program, and for each person within any one program. Studies of special preparation programs carried on in many high schools show various results averaging about 10 points for the verbal section and 15 points for the mathematical over and above the average increases that would otherwise be expected. In other programs, results have ranged from no improvement in scores to average gains of 25-30 points for particular groups of students or particular programs. Recent studies of commercial coaching have shown a similar range of results. You should satisfy yourself that the results of a special program or course are likely to make a difference in relation to your college admissions plans.
- Generally, the soundest preparation for the SAT is to study widely with emphasis on academic courses and extensive outside reading. Since SAT score increases of 20-30 points result from about three additional questions answered correctly, your own independent study in addition to regular academic course work could result in some increase in your scores.

Testing on Campus

ATP tests are available for institutional use outside of the national testing schedule. Colleges and universities can administer the SAT, TSWE, and Achievement Tests on campus to applicants who have not previously taken these tests.

Some colleges that need to know an applicant's scores immediately for admission or placement purposes have the option of scoring the answer sheets on campus. For further information, write to Multiple Assessment Programs and Services, The College Board, CN 6725, Princeton, NJ 08541-6725.

Score Reports for Students: The College Planning Report

The two-page College Planning Report includes the student's test scores, information given by the student on the SDQ, and information provided by the colleges to which the student is having scores sent.

The back of the report contains information on the scoring process and the meaning of scores and percentiles. Accompanying the report is a booklet, *Using Your College Planning Report*, which explains the information received by colleges and how it is used. It also explains how students can use the report to review their college selections.

The College Planning Report

The numbered sections below and on page 9 refer to parts of the sample College Planning Report (pages 10 and 11) for Margaret Wright, a fictitious student. The sample report has corresponding numbers to indicate the part of the report being explained in each of the following sections.

① Identification Information

Much of the information in this section — particularly sex, date of birth, and social security number — is used to retrieve Margaret's data from ATP files, which are stored for the College Board at Educational Testing Service. Submission of her social security number is optional, but it will be used to help identify her record and add scores if she takes ATP tests at another time. It may also help her high school and the colleges that receive her scores to match her record to their files.

② Test Scores

This section shows, for the most recent administration, Margaret's SAT-verbal and SAT-mathematical scores, reported both as specific numbers and as score ranges representing one standard error of measurement (SEM) above and below her numerical scores. (See page 19 for a discussion of SEM.) The TSWE score, but no score range, is also reported here. If Margaret had taken one or more Achievement Tests instead, those scores and score ranges would have been reported here.

Score Ranges. The SEM, rounded to the nearest 10 points, is 30 for Margaret's SAT-verbal score and 40 for her SAT-mathematical score. The score ranges are thus 450-510 for her verbal score and 460-540 for her mathematical score. The score range (or the SEM) is determined by the precision of the test, which is greater for some scores than others. For the SAT, most rounded SEMs will be about 30 points for verbal scores and about 40 points for mathematical scores. Some SEMs will be smaller, particularly for high scores. Presenting the score as a range helps to illustrate that the SAT score gives an approximation rather than a precise measure of ability.

Percentiles. Margaret's report also includes percentile ranks that show the relationship of her scores to the scores of others in each of three reference groups. The percentile rank tells what percentage of that group obtained scores lower than Margaret's. The percentiles section of the score report compares a student's scores with the following reference groups.

- College-bound seniors (national) — all students in the 1985 graduating class who took the SAT or the Achievement Tests at any time while in high school. This reference group includes only the students in a given year's graduating class, and only the most recent SAT and Achievement Test scores for each student are counted.

- College-bound seniors (state) — all students in the state in which Margaret attends high school who were in the 1985 graduating class and took the SAT. (State percentiles are not given for Achievement Tests.)
- National high school sample — a probability sample of all high school students in the nation, based on a special administration of the PSAT/NMSQT in October 1983. (See Appendix A, Table 10.) Students in the sample were not limited to those considering college or planning to take the SAT.

The college-bound seniors percentile ranks used on score reports are updated annually to allow comparisons with the most recent groups of students. Comparisons of this year's percentile ranks with those for previous years can be made by referring to the annual *College-Bound Seniors* series.

Margaret's SAT-verbal score of 480 puts her at the 67th percentile among the college-bound seniors in the nation (see Table 10 on page 22), at the 54th percentile among college-bound seniors in her state, and at the 83rd percentile among all students in the national high school sample who may or may not have enrolled in college.

Margaret's SAT-mathematical score of 500 puts her in the 57th percentile among college-bound seniors in the nation, at the 40th percentile among college-bound seniors in her state, and at the 78th percentile among all students in the national high school sample.

Table 11, page 23, shows the percentile ranks of SAT-mathematical scores for men and women separately. Note that the percentile ranks on the score report are determined from data that combine scores for men and women.

❸ Summary of Test Scores

Because Margaret took the SAT and Achievement Tests during her junior year, the third section of her report shows not only her current SAT-verbal and SAT-mathematical scores, including her verbal subscores and TSWE scores, but also scores for the SAT she took in May 1986 and for three Achievement Tests taken in June 1986. Results for up to six SAT and six Achievement Test administrations may be shown.

Margaret took the SAT the first time as a junior (in May) and scored lower on both the verbal and the mathematical sections than she did when she took the test again the following November. Her new scores, however, are both within the SEM. (See Tables 8 and 9, page 21, for a student's chances of equaling, exceeding, or decreasing scores.)

Scores from the English Composition Test with Essay are distinguished on score reports by the notation ES, but all students receive scores on the 200 to 800 scale, regardless of whether the form contains all multiple-choice questions or a combination of multiple-choice and essay questions. Scores on the two forms can be compared directly and interchangeably. Subscores are not reported for the essay part of the December test.

Specific percentiles for Margaret's scores on the Achievement Tests appeared on the report she received immediately after the June administration. Achievement percentiles appear in Table 13 on page 24.

Note: If an asterisk appears next to a test date in section 3, a message about the scores for that test date is printed immediately under the section.

❹ Educational Background

This section, compiled from information Margaret reported on her SDQ in September 1986 (questions 1-11), shows how many years she studied in each of the academic areas listed, including the arts and music; whether the courses were honors (including advanced placement or accelerated courses); her average grades; and the curriculum covered. It also indicates her report of her grade-point average and class rank.

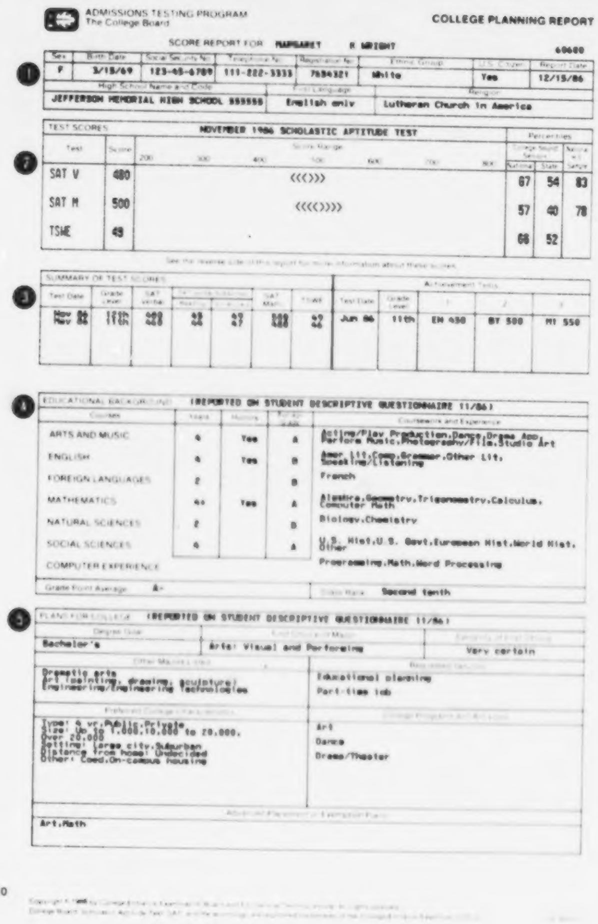
Sections 4 and 5 of the College Planning Report allow Margaret to check that the information submitted on her SDQ is accurate and up to date.

❺ Plans for College

This section, also compiled from the SDQ, shows Margaret's degree goal, first choice of major, and the degree of certainty of her first choice (SDQ questions 20-22). She also could have listed up to four other options for her major (SDQ questions 23-26). Under Requested Services, a student can indicate interest in education and career counseling and developmental academic programs (question 30). Under Preferred College Characteristics, a student can indicate such preferences as location, size, and religious affiliation (questions 14-19). Under College Programs and Activities the student can indicate programs and extracurricular activities that may be of particular interest in college (question 31). Margaret's advanced placement or exemption plans are also indicated (question 29).

❻ Colleges and Scholarship Programs That Received a Score Report

This section provides information on the colleges and scholarship programs to which Margaret has designated that scores be sent, up to a total of eight. Application and financial aid deadlines, address and telephone, and the colleges' criteria for admissions decisions in order of importance — high school record, test scores, extracurricular activities, and so forth — are indicated. This section helps Margaret to keep current with the college application process and allows her to consider her qualifications compared with admissions priorities stated by the colleges she is interested in.



ADMISSIONS TESTING PROGRAM
The College Board

COLLEGE PLANNING REPORT

SCORE REPORT FOR MARGARET R WRIGHT 60680

①

Sex	Birth Date	Social Security No.	Telephone No.	Registration No.	Ethnic Group	U.S. Citizen	Report Date
F	3/15/69	123-45-6789	111-222-3333	7654321	White	Yes	12/15/86

High School Name and Code	First Language	Religion
JEFFERSON MEMORIAL HIGH SCHOOL 555555	English only	Lutheran Church in America

②

TEST SCORES — NOVEMBER 1986 SCHOLASTIC APTITUDE TEST

Test	Score	Score Range (200 300 400 500 600 700 800)	Percentiles: College-Bound Seniors National	Percentiles: College-Bound Seniors State	Percentiles: National H.S. Sample
SAT V	480	<<<>>>	67	54	83
SAT M	500	<<<<>>>>	57	40	78
TSWE	49		68	52	

See the reverse side of this report for more information about these scores.

③

SUMMARY OF TEST SCORES

Test Date	Grade Level	SAT Verbal	SAT Verbal Subscores: Reading	SAT Verbal Subscores: Vocabulary	SAT Math	TSWE
Nov 86 Nov 85	12th 11th	480 420	48 44	49 47	500 480	49 46

Achievement Tests: Test Date	Grade Level	1	2	3
Jun 86	11th	EN 450	BY 500	M1 550

④

EDUCATIONAL BACKGROUND (REPORTED ON STUDENT DESCRIPTIVE QUESTIONNAIRE 11/86)

Courses	Years	Honors	Average Grade	Coursework and Experience
ARTS AND MUSIC	4	Yes	A	Acting/Play Production,Dance,Drama App, Perform Music,Photography/Film,Studio Art
ENGLISH	4	Yes	B	Amer. Lit,Comp,Grammar,Other Lit, Speaking/Listening
FOREIGN LANGUAGES	2		B	French
MATHEMATICS	4+	Yes	A	Algebra,Geometry,Trigonometry,Calculus, Computer Math
NATURAL SCIENCES	2		B	Biology,Chemistry
SOCIAL SCIENCES	4		A	U.S. Hist,U.S. Govt,European Hist,World Hist, Other
COMPUTER EXPERIENCE				Programming,Math,Word Processing

Grade Point Average A+ Class Rank Second tenth

⑤

PLANS FOR COLLEGE (REPORTED ON STUDENT DESCRIPTIVE QUESTIONNAIRE 11/86)

Degree Goal	First Choice of Major	Certainty of First Choice
Bachelor's	Arts: Visual and Performing	Very certain

Other Majors Liked	Requested Services
Dramatic arts Art (painting, drawing, sculpture) Engineering/Engineering Technologies	Educational planning Part-time job

Preferred College Characteristics	College Programs and Activities
Type: 4 yr,Public,Private Size: Up to 1,000,10,000 to 20,000, Over 20,000 Setting: Large city,Suburban Distance from home: Undecided Other: Coed,On-campus housing	Art Dance Drama/Theater

Advanced Placement or Exemption Plans

Art,Math

10

Colleges and Scholarship Programs That Received a Score Report

FOR THE NOVEMBER 1986 TEST ADMINISTRATION

A score report has been sent to the colleges and scholarship programs listed below. The information about the colleges is from *The College Handbook*. For more information about these and other schools, consult the *Handbook* or other materials available in your high school or library and talk with your counselor. Contact the colleges for application materials and additional information.

If you want to have your scores sent to other colleges and scholarship programs, complete an Additional Report Request Form. You received one of these forms with your Admission Ticket. Your high school counselor has additional forms.

[illegible]
City College of Art
[illegible] Street
[illegible] 90512
(213) 123-4567

BASIS FOR ADMISSION DECISION: School achievement record is very important. Interview, school and community activities and recommendations are considered.

ADMISSION APPLICATION DEADLINE: Closing date is August 17. Notification is on a continuous rolling basis.

FINANCIAL AID APPLICATION DEADLINE: No closing date; priority date is March 1. Notification is on a rolling basis starting on April 1.

[illegible]
St. Michael's College
[illegible]
[illegible] Ohio [illegible]
[illegible]

BASIS FOR ADMISSION DECISION: School achievement record and test scores are very important. School and community activities and recommendations are important. Interview is considered. Achievement Tests are required of all applicants.

ADMISSION APPLICATION DEADLINE: Closing date is February 1. Notification begins on or about April 15.

FINANCIAL AID APPLICATION DEADLINE: Closing date is February 1. Notification date is April 1.

[illegible]
State University
80 Central Street
Center City, Texas 34567
(901) 678-9534

BASIS FOR ADMISSION DECISION: School achievement record, test scores and recommendations are important. Interview and school community activities are considered.

ADMISSION APPLICATION DEADLINE: Closing date is August 1. Applications received by April 1 are given priority. Notification is on a continual rolling basis.

FINANCIAL AID APPLICATION DEADLINE: Closing date is April 1. Notification is on a rolling basis beginning January 30.

1920
Alma Mater
645 Southwest Street
Deerfield, New York 10310
(904) 272-6243

BASIS FOR ADMISSION DECISION: School achievement record and test scores are very important. School and community activities and recommendations are important. Interview and school community activities are considered.

ADMISSION APPLICATION DEADLINE: Closing date is January 15. Notification date is April 15.

FINANCIAL AID APPLICATION DEADLINE: Closing date is April 1. Notification date is April 10.

MARGARET K WRIGHT
1234 TIGERLILY LANE
CHICAGO IL 60600

11

Score Reports for High Schools: The College Counseling Report

The score report for high schools (illustrated on page 13) contains, on a single 7 by 11 inch form, most of the information necessary for a college counseling session. Used with a student's College Planning Report, it can help the counselor increase the student's awareness of the range of educational opportunities available. After the student leaves the school, the report remains a record of information for research and statistical reports.

The College Counseling Report contains most of the information found in the College Planning Report, with some variations in formatting. (Note that the College Counseling Report includes the student's test scores only from the most recent administration.) For a discussion of the sections on identification information, test scores, educational background, and college plans, see pages 8 and 9. Section A, which is unique to the College Counseling Report, lists the colleges and scholarship programs, up to a total of eight, to which the student sent scores from the most recent testing.

Using the Report

A comparison of her objective evaluations and her aspirations, as listed on the report, can help the counselor probe Margaret's reasons for her stated educational choices and, if necessary, lead her to some alteration in her thinking. With the report's help, the counselor can ask: Is she aware of her potential? Is she ignoring special talents or interests? Has her high school program adequately prepared her for the college major she plans to pursue?

The counselor can also determine whether Margaret's interests and preferences are reflected in the colleges to which she is sending her scores. Do they reflect the size, location, or religious affiliation she has indicated? Do they offer the types of academic and extracurricular programs that her high school career reflects?

Finally, the report can help the counselor monitor whether Margaret has had score reports sent to the colleges that interest her.

Explaining and Using Score Ranges

Her counselor is in the best position to help Margaret and her family understand the meaning and limitations of test scores. The concept that an SAT or Achievement Test score can only approximately evaluate Margaret's ability is difficult to grasp with a specific numerical score. The visualization of the standard error of measurement—score ranges—along with the explanation on the back of the College Planning Report, should help.

Although the precise score is not an absolute representation of Margaret's ability, the range around this score tends to be a very good measure. Unless there is a compelling reason to think that Margaret had some special problem with a given test, it is unlikely that her score would fall much outside the original score range in a few months' time. Counselors can use this knowledge in starting early college planning with junior test takers, and in advising about retesting, especially if the student has taken a test several times with similar results.

Additional Counseling Materials

As an adjunct to the College Planning and College Counseling Reports, Margaret should be encouraged to avail herself of the many services and publications designed to help her and her family plan for college. Materials produced by the College Board include *The College Handbook* and *Index of Majors*, *ScoreSense** and *College Explorer** (microcomputer programs), audiovisual kits, and numerous publications. For a complete listing, consult the 1986-87 catalog available from College Board ATP CN 6212, Princeton, NJ 08541-6212.

Score Reports for Colleges: The College Admissions and Advising Report

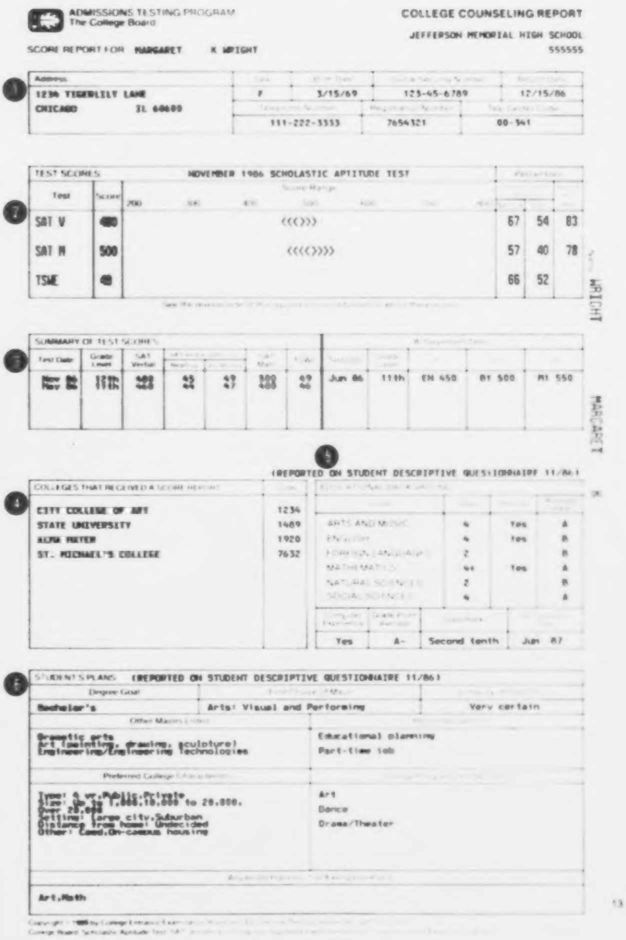
The score report for colleges contains a wealth of information about potential candidates, which can be used before and during the application process and, after enrollment, for placement and advising. The name "College Admissions and Advising Report" reflects its use as more than a source for test scores.

The two-page form, which folds into a standard file with the student's name running across the top, includes both scores and student descriptive information compiled from the SDQ. The other colleges, if any, to which the student had score reports sent are not listed. A sample report is on pages 14 and 15.

The College Admissions and Advising Report

● Identification Information

Sex, date of birth, and social security number (optional) identify student records. If students request that updated reports be sent to colleges at times other than scheduled release dates, comparing the report date, date of SDQ, and the test dates helps to determine which information is current. The telephone number enables admissions officers to communicate quickly with a student for recruitment purposes or for follow-up on an incomplete application. The state and county of residence provide helpful information if the college wants more geographic diversity or if the state or the county is one in which the college intends to conduct greater recruitment activity. A report showing that the student's address differs from the student's legal resi-



ADMISSIONS TESTING PROGRAM
The College Board

COLLEGE COUNSELING REPORT

JEFFERSON MEMORIAL HIGH SCHOOL
555555

SCORE REPORT FOR **MARGARET K WRIGHT**

①

Address	Sex	Birth Date	Social Security Number	Report Date
1234 TIGERLILY LANE CHICAGO IL 60600	F	3/15/69	123-45-6789	12/15/86

Telephone Number	Registration Number	Test Center Code
111-222-3333	7654321	00-341

②

TEST SCORES — NOVEMBER 1986 SCHOLASTIC APTITUDE TEST

Test	Score	Score Range 200 300 400 500 600 700 800	Percentiles		
SAT V	[illegible]	<<<>>>	67	54	83
SAT M	500	<<<<>>>>	57	40	78
TSWE	[illegible]		66	52	

See the reverse side of this report for more information about these scores.

③

SUMMARY OF TEST SCORES

Test Date	Grade Level	SAT Verbal	SAT Verbal Subscores: Reading	Vocabulary	SAT Math	TSWE
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]

Achievement Tests

Test Date	Grade Level			
Jun 86	11th	EN 450	BY 500	M1 550

④

COLLEGES THAT RECEIVED A SCORE REPORT	Code
CITY COLLEGE OF ART	1234
STATE UNIVERSITY	1489
ALMA MATER	1920
ST. MICHAEL'S COLLEGE	7632

⑤ (REPORTED ON STUDENT DESCRIPTIVE QUESTIONNAIRE 11/86)

EDUCATIONAL BACKGROUND

ARTS AND MUSIC	4	Yes	A
ENGLISH	4	Yes	B
FOREIGN LANGUAGES	2		B
MATHEMATICS	4+	Yes	A
NATURAL SCIENCES	2		B
SOCIAL SCIENCES	4		A

Computer Experience	Grade Point Average	Class Rank	
Yes	A-	Second tenth	Jun 87

⑥ STUDENT'S PLANS (REPORTED ON STUDENT DESCRIPTIVE QUESTIONNAIRE 11/86)

Degree Goal	First Choice of Major	
Bachelor's	Arts: Visual and Performing	Very certain

Other Majors Listed

Dramatic arts
Art (painting, drawing, sculpture)
Engineering/Engineering Technologies

Educational planning
Part-time job

Preferred College Characteristics

Type: 4 yr,Public,Private
Size: Up to 1,000,10,000 to 20,000,
Over 20,000
Setting: Large city,Suburban
Distance from home: Undecided
Other: Coed,On-campus housing

Art
Dance
Drama/Theater

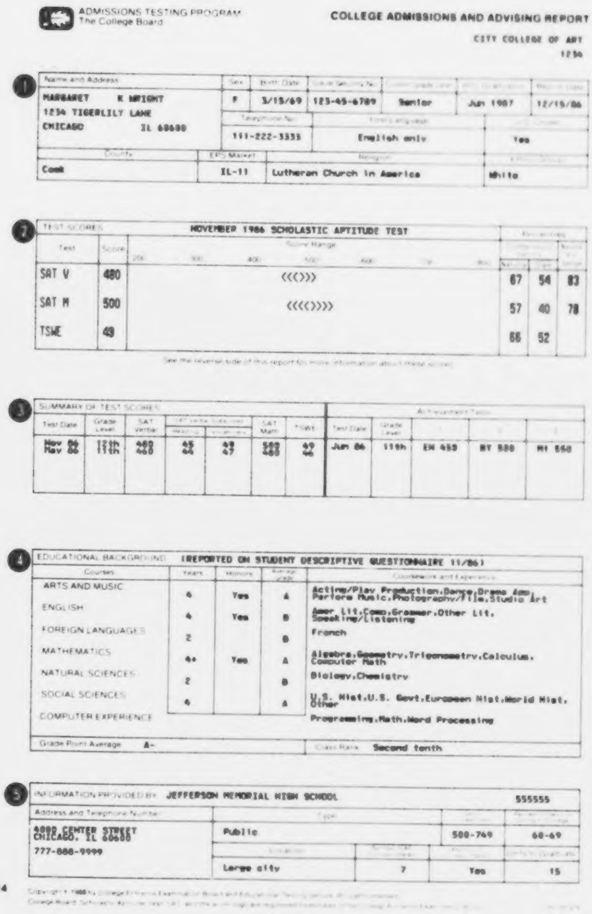
Advanced Placement

Art,Math

Copyright © 1986 by College Entrance Examination Board. [illegible]
College Board, Scholastic Aptitude Test, SAT, [illegible]

WRIGHT
MARGARET

13



ADMISSIONS TESTING PROGRAM
The College Board

COLLEGE ADMISSIONS AND ADVISING REPORT

CITY COLLEGE OF ART
1234

①

Name and Address	Sex	Birth Date	Social Security No.	[illegible]	[illegible]	[illegible]
MARGARET K WRIGHT 1234 TIGERLILY LANE CHICAGO IL 60600	F	3/15/69	123-45-6789	Senior	Jun 1987	12/15/86

Telephone No.	First Language	[illegible]
111-222-3333	English only	Yes

County	EPS Market	Religion	[illegible]
Cook	IL-11	Lutheran Church in America	White

②

TEST SCORES — **NOVEMBER 1986 SCHOLASTIC APTITUDE TEST**

Test	Score	Score Range 200 300 400 500 600 700 800	Percentiles [illegible] National	[illegible]	[illegible]
SAT V	480	<<<>>>	67	54	83
SAT M	500	<<<<>>>>	57	40	78
TSWE	49		66	52	

See the reverse side of this report for more information about these scores.

③

SUMMARY OF TEST SCORES

Test Date	Grade Level	SAT Verbal	SAT Verbal Subscores Reading	Vocabulary	SAT Math	TSWE
Nov 86 Nov 85	12th 11th	480 460	45 44	49 47	500 480	49 46

Achievement Tests

Test Date	Grade Level	1	2	3
Jun 86	11th	EN 450	BY 500	MI 560

④

EDUCATIONAL BACKGROUND **(REPORTED ON STUDENT DESCRIPTIVE QUESTIONNAIRE 11/86)**

Courses	Years	Honors	Average Grade	Coursework and Experience
ARTS AND MUSIC	4	Yes	A	Acting/Play Production,Dance,Drama App, Perform Music,Photography/Film,Studio Art
ENGLISH	4	Yes	B	Amer Lit,Comp,Grammar,Other Lit, Speaking/Listening
FOREIGN LANGUAGES	2		B	French
MATHEMATICS	4+	Yes	A	Algebra,Geometry,Trigonometry,Calculus, Computer Math
NATURAL SCIENCES	2		B	Biology,Chemistry
SOCIAL SCIENCES	4		A	U.S. Hist,U.S. Govt,European Hist,World Hist, Other
COMPUTER EXPERIENCE				Programming,Math,Word Processing

Grade Point Average A-
Class Rank Second tenth

⑤

INFORMATION PROVIDED BY **JEFFERSON MEMORIAL HIGH SCHOOL** 555555

Address and Telephone Number	Type	[illegible]	[illegible]
4000 CENTER STREET CHICAGO, IL 60600 777-888-9999	Public	500-749	60-69

Location	[illegible]	[illegible]	[illegible]
Large city	7	Yes	15

14

Copyright © 1986 by College Entrance Examination Board and Educational Testing Service. All rights reserved.
College Board, Scholastic Aptitude Test, SAT, and the acorn logo are registered trademarks of the College Entrance Examination Board.

(REPORTED ON STUDENT DESCRIPTIVE QUESTIONNAIRE 11/86)

E

HIGH SCHOOL AND COMMUNITY ACTIVITIES	9th 10th 11th 12th
Academic honor society	XXXXXXXXXXXXXX
Art activity	XXXXXXXXXXXXXXXXXXXXXXXXXXXXX OFFICER, AWARD
Dance activity	XXXXXXXXXXXXXX
Religious activity or organization	XXXXXXXXXXXXXXXXXXXXXXXXXXXXX
Theater activity	XXXXXXXXXXXXXX AWARD
Part-time job	XXXXXXXXXXXXXX

F

SPORTS (REPORTED ON STUDENT DESCRIPTIVE QUESTIONNAIRE 11/86)

Basketball Field hockey Softball	Squash Tennis

I

STUDENT'S PLANS (REPORTED ON STUDENT DESCRIPTIVE QUESTIONNAIRE 11/86)

Degree Goal	First Choice of Major	Certainty of Choice
Bachelor's	Arts: Visual and Performing	Very certain

Other Majors Listed	Assistance Desired
Dramatic arts Art (painting, drawing, sculpture) Engineering/Engineering Technologies	Educational planning Part-time job

Preferred College Characteristics	College Programs and Activities
Type: 4 yr,Public,Private Size: Up to 1,000;10,000 to 20,000, Over 20,000 Setting: Large city,Suburban Distance from home: Undecided Other: Coed,On-campus housing	Art Dance Drama/Theater

Advanced Placement or Exemption Plans
Art,Math

WRIGHT MARGARET K F 03-15-69 123-45-6789 dence (shown in the county section) may be an early indication that the student needs to complete documents to establish residency for certain public institutions. Citizenship status can be useful for identifying students eligible for government-sponsored financial opportunities, such as loans, grants, and work study programs, or for requesting additional documentation. A student's religious affiliation or preference and ethnic identity, which are indicated only if requested by the college on the ATP Score Report Options order form, may be used in specialized recruitment. For example, colleges may wish to inform students of the availability of special interest groups on campus.

Test Scores (see College Planning Report)

Summary of Test Scores (see College Planning Report)

Educational Background

The self-reported items in this section of the report are good indicators of a potential for college work. Grades, class rank, honors courses, and expected years of study in certain subjects all reflect the student's interest in learning and response to learning opportunities. There is also a section with specific information on the student's course work and experience. For example, for some institutions or courses of study, it might be essential to know that a student took calculus, not just four years of mathematics.

Honors courses may be considered for placement or credit, or as examples of motivation. Admissions officers may give extra weight to good grades in honors courses, because they probably represent a considerable level of achievement. It is also useful for admissions officers to know about the school the student attended. If, for example, a school offers a number of Advanced Placement and honors courses and a high percentage of seniors attend college, lower grades and class rank might be acceptable.

The total high school program and grades can be considered in conjunction with test scores to determine whether the student's intended majors and educational objectives are realistic. If a student's grades, scores, and high school program demonstrate ability in mathematics and science, for example, is the designated career one in which these abilities will be important? Has the student applied to colleges where training in these areas is available? Does the student have the subject matter depth necessary for the career or major indicated? On the other hand, if the student is undecided about a career or a major, the college may want to inform the student of academic and career opportunities that are compatible with the self-reported interests and grades.

High School Information

The information reported in this section is provided by the high schools.

High School and Community Activities

Margaret's extracurricular experiences — interests, activities, out-of-school learning, community service, and the like — are reported here, with graphics that show how long Margaret has been involved and how current her interests are. If she received honors or served as an officer, this is also indicated.

Sports and Student's Plans

Student responses should be shared with campus groups that might wish to forward appropriate information. Students revealing an interest in sports might welcome a schedule of intramural athletic events. Students expressing interest in participating in other activities could be sent copies of the college newspaper, lists of clubs, programs of cultural or religious activities, or social schedules, as seems appropriate.

Advanced placement or exemption plans can be looked at to see if they are consistent with the student's high school grade in the same subject, to plan for curriculum and staff, or to indicate possible receipt of separately reported Advanced Placement (AP) Program or College-Level Examination Program (CLEP) scores. In planning recruitment activity, admissions officers might want to reexamine how much credit the college offers for examination and how effectively prospective applicants are informed of these opportunities. Perhaps more students would be attracted if they could begin studies at higher levels or take more accelerated courses.

Using the Report as a Marketing Tool

Students can send up to four reports to colleges as part of their basic test fee. The information about students on the reports can help colleges focus their marketing and recruitment activities. First, it enables a college to develop a valuable list of prospects made up of students who not only are aware of the college, but have demonstrated a certain level of interest.

Then, by matching a few student characteristics identified on the report with high-priority marketing interests of the college, each institution can develop key lists for recruitment. The market designation from the Enrollment Planning Service is included in the identification block. The EPS market can be used in conjunction with Enrollment Planning Service data to classify students who submit SAT scores into EPS markets for follow-up and evaluation of recruiting. For instance, the information on the reports will allow colleges to identify quickly students seeking special academic programs, those looking for colleges of a certain size or location, or those with certain extracurricular experiences or high school academic records.

Colleges who receive score reports in tape format will be able to use their computers to separate students by much more detailed criteria.

Using the Report Before Enrollment

Admissions officers can use the extensive information on the report to assess whether the college offers what the student is seeking and whether there is a reasonable chance of admission. Recruitment of potential applicants can be more effective and efficient if students are provided with information that is tailored to their plans, interests, and previous preparation as revealed on their score reports.

In the admissions process, the ATP report — often with a copy of the high school transcript, the completed college application form, letters of recommendation, and personal interview evaluations — is part of the admissions file, the basis for the decision whether or not to admit the student to the institution. Although the high school record may play the strongest role in this decision, ATP scores are particularly useful because they provide a common yardstick of academic potential that is independent of the student's particular curriculum, high school, or region.

Using the Report After Enrollment

After admission, the College Admissions and Advising Report helps in making course placement assignments through the use of SAT, TSWE, and Achievement Test scores and the responses to SDQ questions on subject matter preparation (such as years of study and courses taken in high school), plans for advanced placement, and needs for special assistance. The report can also acquaint college personnel with the characteristics and interests of the student they will be counseling.

Data valuable for planning that involves estimating faculty work loads and the demand on physical facilities can be extracted from the report. Knowing students' housing preferences can help the housing office gauge space needs for the coming year.

Data in the Requested Services section give an early indication of the kinds of services that incoming freshmen think they will need. Department heads and special services personnel might welcome the information for planning for curriculum and staff, establishing reading skills or other developmental centers, determining which counseling areas should be strengthened, and distributing available work-study opportunities.

The Reliability of Self-Reported Information

In making decisions based in part on self-reported information, colleges will want to know how reliable such data are. Evidence shows that student-reported information is often as valid for individual educational decisions as information gathered in more expensive ways. If, as in the SDQ, questions are carefully worded, deal with matters that are relatively recent occurrences, pertain to current concerns and interests, and can be verified, answers to them may be used with a degree of assurance. (See *Using Self-Reports to Predict Student Performance*.)

Understanding College Board Scores

College Board scores for the SAT and the Achievement Tests are reported on a scale of 200 to 800. The choice of any score scale becomes meaningful only as data are compiled from the scores of various groups of students taking the test. Users learn to understand and appreciate the meaning of a score of 430 in the same way that they have learned to understand and appreciate the meaning of, say, 14 inches, a process that is possible only if the measuring units remain constant. In the early years of the SAT, the test was rescaled each year so as to provide an average score of 500. Since 1941, however, a constant scale has been used that is maintained through a process known as equating. Scores on each new form of the SAT or the Achievement Tests are calibrated against prior forms. As a result, different forms of a test such as the SAT yield scores on the same 200 to 800 scale, thereby enabling the user to compare scores of students who take the test at different times. For example, within the limits of the equating methods employed, a verbal score of 430 on the SAT today represents the same level of developed verbal ability as it did several years ago. Therefore, students' scores can be compared from one administration or one year to another.

In addition to calibration against prior forms, Achievement Test scores are maintained by periodic rescaling studies in order to make scores from different tests roughly comparable. Because rescaling may affect the placement of the tests on the scale, some year-to-year differences in scores may be due to rescaling as well as performance. Achievement Tests have not been rescaled since 1980.

Equating procedures are under continuous review. A 1976 study verified that the SAT scale had not shifted substantially between 1963 and 1973 and that the score declines reported nationally could not be attributed to a drift in the score scale.

Separate verbal and mathematical scores are reported for the SAT on the 200 to 800 scale. SAT verbal subscores for reading comprehension and vocabulary are reported on a 20 to 80 scale. (Note that averaging the two subscores and multiplying by 10 does not result in the SAT verbal score.)

Reading comprehension subscores are obtained from SAT reading passages and sentence completion questions; vocabulary subscores come from the analogy and antonym questions. Because reading comprehension and vocabulary are closely related, the difference between the reading comprehension subscore and the vocabulary subscore for a student has low reliability. Only when the difference between the two is as great as 9 points can one be certain that there is a genuine difference in the abilities being measured.

Scores on the TSWE are placed on a 20 to 80 scale; however, because the TSWE is not intended to distinguish among students whose command of standard written English is considerably better than average, the maximum reported score is 60+.

Raw and Scaled Scores

Before students' scores can be placed on the College Board scale, "raw scores" must be obtained. Each correct answer receives one point and no points are assigned to omitted questions. A correction for guessing is then applied. For questions with five answer choices, one-fourth of a point is subtracted for each incorrect response; one-third of a point is subtracted for incorrect responses to questions with four answer choices. For example, if, on a test of 85 questions with 5-choice responses, a student has 44 right, 32 wrong, and 9 omitted, the resulting raw score is determined as follows:

44 right − ¼(32 wrong) = 36

Raw scores for each new form of a test are placed on the College Board scale through the equating process. Table 1 shows the relationship of SAT and TSWE raw scores to the College Board reporting scale (scaled scores). Although the SAT is constructed to meet precise difficulty-level specifications, there is inevitably a small amount of variation in difficulty from one edition to the next. Therefore, a lower raw score is needed on a more difficult form to obtain a given scaled score than is needed to get that same score on an easier form. For example, in Table 1, a raw score of 35 on a more difficult form of the verbal sections of the SAT will produce a scaled score of 430, whereas an easier form will result in a scaled score of 420.

A scaled score of 200 does not necessarily stand for a minimum raw score; it is the lowest score reported. The scaled score of 800 does not necessarily stand for a perfect raw score; it is the highest score reported.

Measurement Characteristics of ATP Scores

Analyses that provide information about the measurement characteristics of ATP tests are performed regularly for each new form. The data obtained from each test analysis provide information about the test's reliability, the difficulty and speededness for the group tested, and the intercorrelation of scores for the test components. Tables 2

BEST COPY AVAILABLE

[illegible]

[illegible]

to 4 provide some measurement characteristics for three recent editions of the SAT and the TSWE. Table 5 provides similar information for the Achievement Tests.

Reliability

The precision of any test score is limited because it represents only a sample of all the possible questions that could be asked, and because people perform at different levels at different times for reasons unrelated to the characteristics of the test itself.

The consistency with which the test measures true performance is expressed as a "reliability coefficient." It indicates the extent to which an individual would achieve the same score on repetition of a test. A reliability coefficient of zero indicates no relationship whatsoever between a student's relative standing within a group on two forms of a test, whereas a reliability coefficient of 1.00 indicates perfect reliability — students within a group rank exactly the same on the two forms.

The reliability data for the SAT scores in Table 2 were derived from item response theory (IRT) estimates of standard errors of measurement (see the next section, Standard Error of Measurement, and Appendix A). The reliability data for the TSWE and the multiple-choice Achievement Tests included in Tables 2 and 5 were obtained using the Kuder-Richardson Formula (20) with the Dressel adaptation for formula-scored tests. Both reliability estimates are influenced by differences in examinee performance due to the sample of questions selected and the degree to which the questions vary in content. The estimates do not take into account day-to-day differences in examinee behavior or differences in administration environment.

The reliability data for the SAT and TSWE are based on statistically representative samples of juniors and seniors taking these tests. The reliability estimates for the verbal and mathematical sections of the SAT, comprising one hour of testing time for each component, are typically about .91. For a typical TSWE, the reliability coefficient is about .89.

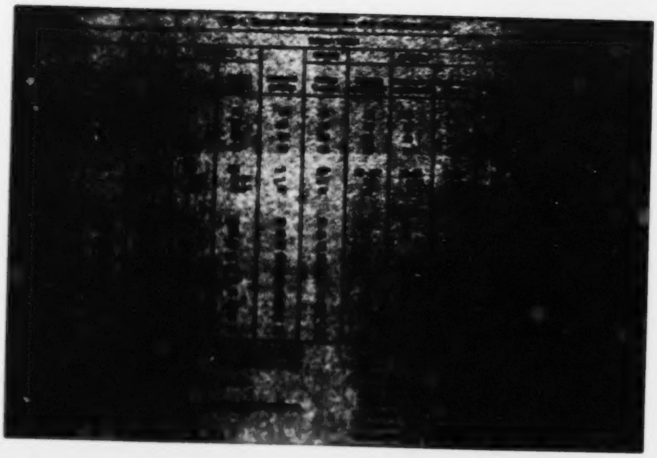
The reliability coefficients shown in Table 5 for the one-hour Achievement Tests range from .86 for Mathematics Level I to .95 for German.

Standard Error of Measurement

The most realistic way to allow for the effects of normal variations in the physical and emotional conditions of the individual, the test setting, or the test content is to interpret scores as ranges rather than as points. The same test or a different version of a test taken on different days would probably result in a slightly different score each time. If a student were to repeat the test an infinite number of times, a number of different scores would probably be obtained, some higher, some lower, but most would tend to cluster about an average value. This average would be the "true score," the score a student would earn if the test could measure ability with perfect reliability. An index of the extent to which students' obtained scores differ from their "true scores" is called the standard error of measurement (SEM). The SEM for a given test can vary at different places on the same scale. For example, the SEM for SAT-verbal scores is approximately 30 points for scores 200 to 670 and about 20 points for scores 680 to 800, with the average SEM being about 30 points. For SAT-mathematical scores the values for the SEM are approximately 30 points for scores 200 to 330, 40 points for scores 340 to 530, 30 points for scores 540 to 700, and 20 points for scores 710 to 800. This results in an average SEM for the SAT-mathematical score of about 35 points. The SEMs reported for each ATP test are, in effect, an average of the SEMs for that test. The SEMs for ATP tests are reported in Tables 2 and 5.

The SEM for the SAT-verbal score is approximately 30 points on the 200 to 800 scale. This means that two-thirds of the students taking the test will obtain scores within 30 points above or 30 points below (one SEM) their true score. For example, if a student has a true score of 430, the chances are about 2 out of 3 that the student will receive an obtained score between 400 and 460 (430 plus or minus 30). The score ranges shown on the score reports illustrate this concept for students, counselors, and admissions officers.

The SEM for the SAT-mathematical score is approximately 35 points on the 200 to 800 scale. For most students this will be rounded to plus or minus 40 points on their score reports. The standard errors of measurement for the SAT-verbal subscores are about 4.4 points on the 20 to 80 scale for reading and about 4.6 points for vocabu-



lary. The standard error of measurement for the TSWE is about 3.6 points. For the all-multiple-choice Achievement Tests, the SEM range is from a low of about 24 points for the Hebrew and Spanish Tests, to a high of about 36 points for the Literature Test.

Standard Error of the Difference

Users of test scores are advised against making fine distinctions between scores. The standard error of the difference, which is reported in Tables 2 and 5, indicates the normal variation to be expected between the scores of two people on the same test, or tests taken at two different times by the same person, due to measurement error alone. Score differences of less than 1.5 times the standard error of the difference have little significance. For example, the standard error of the difference for the SAT-mathematical scores is about 48. Only when scores differ by more than 72 points (48 x 1.5) can there be reasonable confidence that the abilities being measured genuinely differ.

Speededness

Detailed test analyses suggest that ATP tests are relatively unspeeded for the majority of the students tested and that most scores would not appreciably improve if more time were allotted.

A test may be considered unspeeded if virtually all of the students taking it complete three-fourths of the questions and 80 percent reach the last question. The percentage completing three-fourths of the test is a more reliable indicator than the percentage completing the final question because the last question is often difficult and students may be omitting rather than not reaching this question.

By the first standard, the test sections of three forms of the SAT and one form of all but three Achievement Tests were slightly speeded for some students. The percentages completing three-fourths of the separately timed SAT and TSWE sections ranged from 97 to 100 (see Table 3). For the Achievement Tests, the range was 96 to 100 (see Table 5).

Another indicator of speed included in Tables 3 and 5 is the average number of questions not reached by each sample of students. For the separately timed SAT-verbal, SAT-mathematical, and TSWE sections, only one to two questions, on the average, were not reached by the representative group of juniors and seniors taking the test. The average number of questions not reached on the Achievement Tests ranged from less than one for German, Latin, and Literature examinees to approximately three for the students taking American History and Social Studies, Hebrew, and Spanish.

Intercorrelation of Components

The intercorrelation of the different components of the SAT and TSWE for a typical form is presented in Table 4. The correlation coefficient between the verbal and mathe-

BEST COPY AVAILABLE

matical scores is approximately .66; between the reading comprehension and vocabulary subscores, about .80. The level of correlation for verbal and mathematical scores suggests that there is some overlap in the information provided by these tests. An even greater degree of overlap is indicated for the verbal subscores, suggesting that only in exceptional cases would students obtain very different subscores. The correlation coefficient between the multiple-choice and essay components of the English Composition Test is .45. The degree of the relationship is limited because there is only one essay question on which to sample writing ability. But even if there were more essay questions, the correlation between the two parts would not be perfect because some unique as well as some common skills or abilities are being measured. This relationship and those noted for the SAT and TSWE are typical of other forms of the ATP tests.

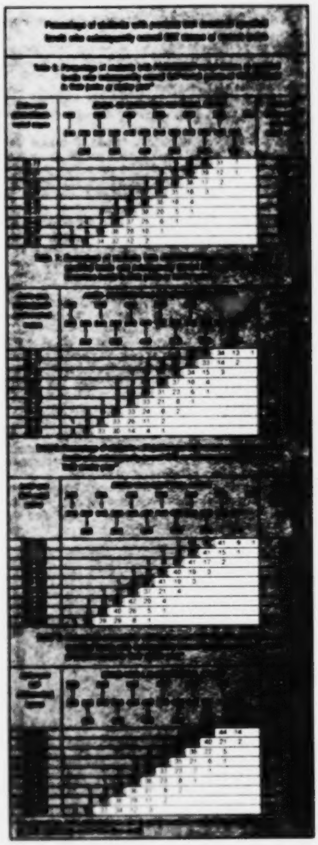
Repeating Tests

When students take tests more than once, their scores usually change. This change may be due to the practice effect, to academic growth, or to other influences. However, the most powerful influence on this change is the imprecision inherent in test scores, which, as noted previously, is indicated by the standard error of the difference when two scores are compared. Thus, score increases or decreases can be as large as 1.5 to 2 standard errors of the difference and still not indicate any real difference in the student's ability.

In the case of the SAT, an appreciable rise is unlikely. On average, students who took the SAT as juniors in spring 1985 and again as seniors in fall 1985 improved their verbal scores by about 15 points and their math scores by about 21 points. Furthermore, for those students whose scores change when they repeat the test, about 65 percent have score increases, while about 35 percent have score decreases. The higher a student's initial scores, the greater the probability that subsequent scores will be lower. The lower the initial scores, the more likely the subsequent ones will be higher. Among students repeating the SAT, about 1 in 20 gains 100 or more points and about 1 in 100 loses 100 or more points.

Tables 6 and 7 show the percentage of students with junior year PSAT/NMSQT scores at various levels who earned SAT scores at various levels. Table 6 refers to verbal scores; Table 7 refers to mathematical scores. For example, the number "3" in the top row of Table 6 means that among the students with junior year PSAT/NMSQT verbal scores of 68 to 72, approximately 3 percent earned SAT verbal scores of 550 to 590 in their junior or senior year. The percentages in each row add up to 100 percent (in some cases, plus or minus 1 percent due to rounding). The column at the right side of Tables 6 and 7 shows the average SAT score for the students with PSAT/NMSQT scores within each specified range.

Tables 6 and 7 show that the students' junior year or senior year SAT scores tend to vary in both directions from their PSAT/NMSQT scores. Students with PSAT/NMSQT scores in a 5-point range (corresponding to a 50-point SAT score range) may earn SAT scores that differ by 200 or more points. This tendency is slightly greater for mathe-



matical scores than for verbal scores.

In interpreting Tables 6 and 7 note that the results are somewhat affected by the bell-shaped distribution of scores on the PSAT/NMSQT. The farther the PSAT/NMSQT score is from the middle of the distribution, the fewer students get that score. For example, among the students with PSAT/NMSQT scores of 68 to 72, there will be many more scores of 68 than 72. Similarly, among those with scores of 28 to 32, there will be many more with 32 than with 28.

Tables 8 and 9 show the percentages of students with junior year SAT scores within various ranges who subsequently earned senior year SAT scores at various levels. Table 8 refers to verbal scores; Table 9 refers to mathematical scores. The interpretation of these tables is similar to that of Tables 6 and 7. For example, the number "1" in the top row of Table 8 means that among the students with junior year SAT-verbal scores of 680 to 720, approximately 1 percent earned SAT-verbal scores of 550 to 590 in their senior year. The column at the right of each table shows the average senior year SAT score for students with junior year SAT scores at each specified level.

The spread of senior year scores for the students in each junior year score category is substantial, though smaller than when the categories are based on PSAT/NMSQT scores. Students with junior year scores in the same 50-point interval may have senior year scores 200 or more points apart. Again, the statistics in Table 8 and 9 reflect effects of the bell-shaped score distribution. The highest category includes many more scores of 680 than of 720; the lowest category includes more scores of 320 than of 280.

At a specific high school, the average change in scores for students who repeat a test may differ substantially from the average change of the national group simply because of the sampling error present in small samples.

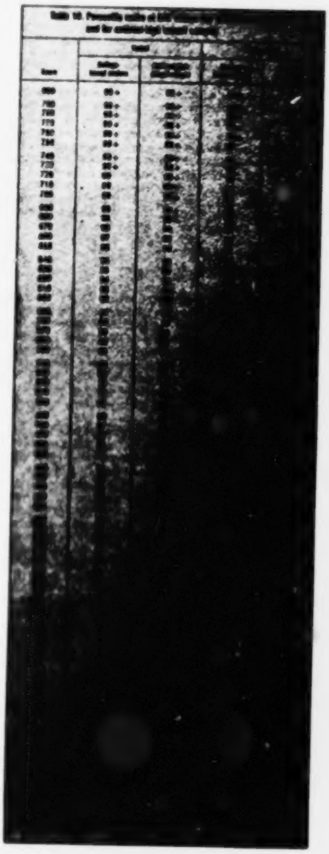
In the case of an Achievement Test, which is designed to measure a student's knowledge of a subject, score increases may result because the student has studied the subject for another semester or two.

Using ATP Scores

The following suggestions, which are far from exhaustive, are intended to stimulate ideas that will yield the greatest benefit from the reported data. (See Appendix C, Guidelines on the Uses of College Board Test Scores and Related Data.)

SAT Scores

When scores or other data are used for selection (to accept a student for admission or to permit entry into a particular course), it is important that they be validated periodically, most appropriately through a validity study to insure that they predict the expected outcome at a level acceptable for the institution's particular purpose. Every three years is a generally accepted standard. A validity study also provides the relative weight that should be given



BEST COPY AVAILABLE

to scores and other data in predicting how a particular student will perform: for example, the weight to be given SAT scores and the high school record to predict grade point average (GPA); the weights to be given Achievement Test scores, SAT scores, and high school record for admissions purposes; and the level of Achievement Test or TSWE scores that best predicts acceptable performance in the particular course.

Some students may be anxious about their SAT scores because they could not answer all or most of the test questions correctly. The data in Table 2 (page 18) may reassure them. For example, on the average, students answer only about half of the SAT-verbal questions correctly, resulting in a score in the 410 to 440 range.

Some students may be discouraged by what they consider low scores. Students tend to evaluate their scores in the belief that the average College Board test score is 500. A review of Table 10 (page 22) with the student may be helpful, pointing out that an SAT-verbal score, for example, of 500 is at the 87th percentile of all high school seniors and that a 370 score approximates the median score. If the student's concern centers on the difference between his or her scores and some other student's, the data in Table 2 on the standard error of the difference can indicate if the difference is likely to be real or simply the result of the imprecision of the testing process. For example, two SAT-mathematical scores would have to differ by more than 72 points (1.5 × the standard of error of the difference) for one to be reasonably confident that the higher score represents more highly developed mathematical aptitude.

Students sometimes repeat the SAT in the hope of improving their score and they wonder which score will be used. Most admissions officers consider all the scores in a student's report. However, some admissions officers prefer to give students credit for their best performances and use the highest scores. The student who takes the SAT two or three times will probably receive at least one score higher than the score of the equally capable student who takes it only once. When admissions officers use only the highest scores, the student who can afford to take the test only once might be at a disadvantage compared with the student who has taken the test more than once.

Some admissions officers prefer to use a student's most recent SAT scores. This choice may be less subject to error of measurement than using the student's highest scores. The most recent scores may better reflect a student's current ability.

Other admissions officers calculate an average of all the student's SAT-verbal scores and an average of all the SAT-mathematical scores. This method may be the most equitable of the three if scores span a short period. It may be helpful to compare SAT scores with the high school record. Unusually high SAT scores and weak high school records may indicate able students who have not applied themselves in high school. Very low scores and strong high school records may indicate students who work hard and achieve through perseverance. Other data such as teacher and counselor recommendations may be needed to assess accurately students' readiness for a certain college.

The booklet *Taking the SAT* contains a sample test, answers, and a table giving percentages of a random sample of students correctly answering each question on the May 1983 SAT and the June 1981 TSWE. The booklet also contains scoring instructions and can be used to gain a clearer idea of what is tested and of what a score represents.

TSWE Scores

Studies have been conducted that demonstrate the effectiveness of using TSWE scores in college placement

With the cooperation of 15 institutions, freshman English course grades and TSWE scores were obtained for over 4,000 students. A comparison of the grades and scores showed a substantial relationship, with over 90 percent of those scoring in the highest score range on the TSWE having an A or B in freshman English courses.

A second study concentrated on the relationship between direct measures of writing ability (essays) and indirect measures (multiple-choice tests) like the TSWE.

Table 14 on page 25 shows the relationship between TSWE scores and the percentages of students writing above-average essays before instruction. Of the students with the highest scores (60+), 85 percent wrote above-average essays at the beginning of an English writing course.

These studies and other data obtained in the administration of the test confirm the appropriateness of the TSWE in terms of difficulty and of discriminating power. The test meets the purpose for which it was designed, that is, to identify students who might benefit from additional or specialized instruction in standard written English. Colleges should develop institution-specific placement applications of the TSWE. (See *Methods of Implementing College Placement and Exemption Programs*.)

Achievement Test Scores

Achievement Tests are curriculum based, but they are independent of particular textbooks or methods of instruction. They are designed to assess outcomes of courses that students have taken recently. If a student completes a biology course in the tenth grade but does not take the Biology Achievement Test in that subject until the twelfth grade, the time lag may place the student at a disadvantage. If a student takes the Achievement Test in Chemistry before completing the course of study, the student is also at a disadvantage. If a student takes the same Achievement Test more than once, the score for the test taken closest to the completion of the course would presumably be more indicative of the highest achievement level attained.

In using an Achievement Test score for placement, consideration should be given to the number of years of study in the subject and the level of courses taken. For example, high school students may have studied a language for different periods of time. Most students who take an Achievement Test in a foreign language choose to do so during the third or fourth year of language study. Others, however, take the test during their second year of study or even their fifth. Candidates who take ATP foreign language examinations are asked to supply certain information about their training and experience in the language. Normative data that provide distributions of test scores earned by students who have taken two, three, and four years of study, useful for evaluating student performance on the basis of years of coursework, can be obtained by writing College Board ATP (see inside front cover).

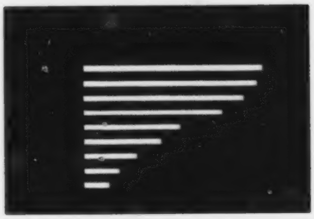
If the difference between the scores earned by two different people exceeds 1.5 times the standard error of the difference for the test (see Table 5 on page 20), one can be reasonably confident that the higher score reveals greater ability or achievement as measured by the test. A difference of fewer than 66 points in the scores of two students on the Biology Achievement Test, for example, should not be considered significant.

The comparison of scores earned by two students on Achievement Tests in different subjects is a more complicated matter. Although every Achievement Test score is reported on the same 200 to 800 scale, individual scores earned on different Achievement Tests are only roughly comparable. It is best to avoid comparing scores earned by different students on different Achievement Tests.

Scores of Students for Whom English Is a Second Language

If English is not the student's first language, exercise judgment in estimating how much a limited facility with the language may have affected grades and test scores. Although no clear-cut pattern holds for students from every part of the world, students from outside the United States generally do better on the mathematical questions of the SAT and on Achievement Tests in mathematics and the sciences than they do on the verbal questions of the SAT, the TSWE, and the Achievement Test in English Composition. Performance on the foreign language Achievement Test specific to the student's native language may reflect factors other than classroom achievement. Although facility in that language is accurately assessed, the Achievement Test score may not be a significant predictor of overall academic performance.

The Test of English as a Foreign Language (TOEFL) was designed to help assess a foreign born student's grasp of English. Performance on the TOEFL may serve to help interpret scores on the TSWE or on the verbal sections of the SAT. For example, if a student's TOEFL scores are low and the score on the verbal sections of the SAT is also low, it may be inferred that performance on the verbal sections of the SAT was probably affected by the student's deficiencies in English. For further information about the relationship between language proficiency as measured by TOEFL and performance on the SAT-verbal sections and the TSWE, see *TOEFL Research Report 3*. For information on TOEFL test dates and center locations, see *TOEFL Test Center Reference List*.



Counselors and admissions officers should also be alert to the problems of students from countries other than the United States who speak excellent English but who may be at a disadvantage because of their unfamiliarity with testing methods in the United States and because ATP tests naturally reflect a United States cultural background.

Scores of Minority Students

The College Board makes every attempt to ensure that test content is as fair as possible to all groups. A sensitivity review committee has drawn up guidelines for the types of minority-relevant content to be included in the SAT. This committee also reviews tests to eliminate questions that depend on words that may have different meanings for various groups. Preliminary review helps to eliminate the inclusion of biased questions into final copies of the test. Statistical analyses are routinely performed to identify questions or types of questions on which performance differs for different groups of students.

Admissions officers have found it a wise policy to ensure that members of minority groups are not excluded on the basis of test scores alone. Other factors such as high school grades, strong motivation, and maturity of purpose can indicate a potential for success.

Validity studies conducted by individual colleges indicate that the test scores usually are very useful in predicting freshman grade point averages for minorities. Each college is encouraged to conduct its own validity study for minority students.

Scores of Students with Handicaps

The message "Nonstandard Administration" on a report and an asterisk next to a test date indicate that the student took the test in a nonstandard administration.

The purpose of special test arrangements is to attempt to minimize the impact of an individual handicap in the test situation so that students can demonstrate their academic ability.

The individual circumstances that require special test arrangements are so diverse that the College Board is not able to provide meaningful interpretive data for scores earned in nonstandard administrations. The total number of students with handicaps who have taken special editions of the SAT is small, and the number who attend any given college is smaller still, so the correlations between test scores and the first year averages of these students have not been established. The norms from standard administrations published in this Guide may be a useful reference, but the usual caution that test scores should be considered only one factor in the assessment of a student's academic potential is especially applicable when the scores of students with handicaps are being interpreted. (See ATP Services for Handicapped Students: *Information for Counselors and Admissions Officers* and *Information for Students with Special Needs*.)

Scores of Adult Students

Scores from tests taken more than five years ago are probably not good indicators of a person's current ability to do college work. A person whose work requires verbal or mathematical abilities or who reads widely or studies independently may be better qualified for college work now than at high school graduation. After five years, it is recommended that students take the tests again rather than rely on the old scores.

Admissions officers may find that ATP score reports received for adult applicants are best evaluated in terms of their most recent experiences. High motivation and the contributions of life experience are important considerations in predicting college performance for adult students.

The mean scores for 31,066 adults (ages 20 and over) who took the SAT during the 1981-82 school year are shown in Table 15. On the average, adults earned lower mathematical scores than did college-bound seniors. However, adults from 25 to 39 years of age earned higher average verbal scores than college-bound seniors. As with college-bound seniors, adult men scored higher on the mathematical portion of the test than did adult women.

Percentiles

Percentiles should not be viewed in isolation but rather as additional information about a student's test scores and what they mean in comparison with various groups of students. For an explanation of percentile ranks and of the reference groups, see pages 8 and 9. The percentile ranks of two different reference groups — for instance, men and women — should not be compared. However, the scaled scores for two different groups on the same test can be compared.

Using Predictive Validity Studies

Predictive validity indicates a test's effectiveness in predicting a student's performance. This, fundamentally, is the purpose of the ATP tests—to serve as predictors of academic performance in college. Such information has proved helpful to admissions officers who can use test scores along with other academic aspects of the secondary school record to predict a student's chances of academic success in college.

If students who do poorly on the test do poorly in college and those who do well on the test do well in college, the test is said to have a high predictive validity. Thus the statistical task is one of measuring the degree of association between the predictors (test scores and high school record) and the criterion (grades in a particular college course, a general freshman average, or a four-year average). This degree of association is expressed as a correlation or validity coefficient whose values range theoretically from −1.00 to +1.00.

Table 16 presents a summary of all validity studies conducted through the Validity Study Service that were designed to predict freshman grade point average, using SAT scores and high school record (rank or average). All studies of the whole freshman class, of male freshmen, of female freshmen, and of freshmen entering any of six selected college curriculums are included. (If a college conducted more than one study of any type, only the most recent one is included.)

For the 685 colleges that studied their whole freshman class, the 90th percentile, median, and 10th percentile validity coefficients were almost the same for the SAT verbal and SAT mathematical scores. That is, the SAT verbal correlation was above .52 for 10 percent of the colleges, between .36 and .52 for 40 percent, between .21 and .36 for 40 percent, and below .21 for 10 percent. The SAT math correlation was above .50 for 10 percent of the colleges, between .35 and .50 for 40 percent, between .20 and .35 for 40 percent, and below .20 for 10 percent.

The validity of high school record is typically somewhat higher than the validity of the optimally weighted combination of SAT scores. For example, for all freshmen, the median correlations for high school record and for the optimally weighted combination of SAT scores were .48 and .42, respectively. For males, they were .45 and .39. For females, they were .50 and .46.

[illegible]

The validity of the optimally weighted combination of high school record and SAT scores is usually higher than that of either the high school record or SAT scores separately. For example, for all freshmen, using the combination of high school record and SAT scores raised the median correlation .13 over SAT scores and .08 over high school record. (Because the median correlations in Table 16 are rounded, the difference appears to be .07.) Although such improvements may seem small, they represent an appreciable increase in the accuracy of academic prediction.

Still greater accuracy in predicting college performance can often be obtained by using an applicant's Achievement Test scores, either individually or averaged, in addition to high school record and SAT scores. The average increase in correlation is between .02 and .03, bringing the total ATP test score increase to .10-.11 over high school record.

The ranges of high school grades and test scores of those who enroll in any given college are smaller than those of all students going to college. Applicants with low grades or scores are often not accepted. In addition, students with either unusually high or low grades or scores for a given college select themselves out by not applying. As the variability among enrolling freshmen on grades and scores is reduced from that which would have been observed if the college and the students had not used grades and scores to decide who will enroll at the college, the correlation coefficients are reduced. In the extreme case, if all students at a college have the same grades or scores, no prediction is possible. Because there is even less variability within a college curriculum than for an entire freshman class, validity coefficients by curriculum are even more restricted.

However, whenever the number of students is sufficiently large, colleges should do separate studies by college curriculum (especially because courses may differ and grades may not be comparable), sex, ethnic group, and other relevant subgroups of students. Separate studies would show whether use of grades and test scores is fair for each subgroup and whether use of separate prediction equations might promote greater equity.

The use of statistical predictions based on the high school record and whatever test scores are available gives what is theoretically the best possible indication of the student's college grade record that can be made from these data. With this information available for a given applicant, the admissions officer is free to devote more time to considering the school's recommendations and other information about the applicant's personal qualifications.

For additional information, see Chapter 8 of *The College Board Technical Handbook for the Scholastic Aptitude Test and Achievement Tests.*

Appendix A. Tables and Their Sources of Data

Table 1. Raw scores converted to scaled scores (page 18)

The ranges for the College Board scores are based on all new editions of the SAT and the TSWE given from January 1982 through June 1985.

Table 2. Sample statistics for three recent forms of the SAT and the TSWE (page 18)

Data in this table were derived from item analyses for representative samples of juniors and seniors taking three recent forms of the SAT and TSWE administered in November 1984, December 1984, and May 1985. The mean scores for the three samples reflect the differences in ability of the three student groups who elect to take the SAT and TSWE at different times of the year.

For the SAT, item response theory (IRT) standard errors of measurement (SEMs) were estimated for each scale score point; weighted averages of these SEMs are reported in Table 2. Reliabilities were computed as one minus the ratio of the average squared SEM to total score variance. For more details, see:

Dorans, N.J. "Approximate IRT formula score and scaled score standard errors of measurement at different ability levels." ETS Statistical Report No. SR-84-118, 1984.

Lord, F.M. *Applications of Item Response Theory to Practical Testing Problems*. Hillsdale, NJ: Lawrence Erlbaum Associates, 1980.

For the TSWE, reliability coefficients, SEMs, and standard errors of the difference were computed by applying the Kuder-Richardson Formula (20), Dressel adaptation for formula-scored tests.

Table 3. Speededness of sections for the SAT and the TSWE (page 19)

Speededness data are based on the same three samples and three test forms described for Table 2. Each section is allotted 30 minutes of testing time and data are provided for each of the separately timed parts of the verbal and mathematical sections of the SAT and of the TSWE.

Table 4. Intercorrelations of SAT components and the TSWE for a typical recent form (page 19)

The coefficients of correlation provided in this table are based on the sample of 1,500 students who took the November 1984 administration described in Table 2. The coefficients are not corrected for the unreliability of the test components.

Table 5. Measurement characteristics of the Achievement Tests (page 20)

Data for this table were obtained from item analyses for random samples of the test-taking populations or based on samples stratified by previous levels of study. (See footnote 1.) The reading reliability estimate for the essay component of the English Composition Test (ECT) is obtained by first computing the correlation coefficient between individual reader scores to obtain an estimate for a single score, and then applying the Spearman-Brown Formula to obtain an estimate for the sum of two reader scores.

Table 6. Percentage of students with PSAT/NMSQT verbal scores at specified levels who subsequently earned SAT verbal scores at various levels in their junior or senior year (page 21)

Data are based on about 891 000 students who took the PSAT/NMSQT as juniors in October 1980 and the SAT either as juniors in the spring of 1981 or as seniors in the fall of 1981.

Table 7. Percentage of students with PSAT/NMSQT mathematical scores at specified levels who subsequently earned SAT mathematical scores at various levels in their junior or senior year (page 21)

Data are based on about 891 000 students who took the PSAT/NMSQT as juniors in October 1980 and the SAT either as juniors in the spring of 1981 or as seniors in the fall of 1981.

Table 8. Percentage of students with junior year SAT verbal scores at specified levels who subsequently earned SAT verbal scores at various levels in their senior year (page 21)

Data are based on about 375 000 students who took the SAT in the spring of 1985 as juniors and in the fall of 1985 as seniors.

Table 9. Percentage of students with junior year SAT mathematical scores at specified levels who subsequently earned SAT mathematical scores at various levels in their senior year (page 21)

Data are based on about 375 000 students who took the SAT in the spring of 1985 as juniors and in the fall of 1985 as seniors.

Table 10. Percentile ranks of SAT scores (for college-bound seniors and for national high school sample) (page 22)

Data for the college-bound seniors columns include 977 000 (472 000 males and 505 000 females) who were in the high school graduating class of 1985 and who took the SAT prior to May of their senior year. This reference group includes only the students in the given high school graduating class and year and only one set of SAT scores per student is counted (the most recent one prior to May of the senior year). This conforms to data published annually in *ATP College-Bound Seniors*. Table 10 is the source for the SAT percentile ranks found on ATP score reports. Percentile ranks for the national high school sample are estimated from a stratified cluster sample of 25 316 juniors who took a special administration of the PSAT/NMSQT in October 1983. The students who were sampled constituted the junior classes of a probability selection of 105 secondary schools, drawn from a national frame containing approximately 26 000 public, parochial, and private schools. The estimation process used the PSAT/NMSQT to SAT growth data described in Tables 6 and 7. The percentile ranks are estimates of those that would be obtained if all juniors and seniors took the SAT. They are not based only on students considering college or those who normally take the SAT.

Table 11. Percentile ranks of SAT mathematical scores for men and women separately (page 23)

See Table 10 for a fuller description of the national high school sample. These percentile ranks are not used on ATP score reports; reported percentile ranks are determined from data that combines scores for men and women as shown in Table 10.

Table 12. Percentile ranks of TSWE scores (for 1985 college-bound seniors) (page 23)

Percentile ranks are based on scores earned by 977 000 high school students who were college-bound seniors in 1985. See the Table 10 entry for a description of the college-bound population.

Table 13. Percentile ranks of Achievement Test scores (for 1985 college-bound seniors) (page 24)

Percentile ranks are based on scores earned by high school students who were college-bound seniors in 1985. See the Table 10 entry for a description of the college-bound population; in reading the description, substitute Achievement Test for SAT. Actual numbers of college-bound seniors electing to take these tests may be found in Table 13.

Table 14. Percentage of students writing above-average essays at the beginning of the first course (page 25)

The data were collected in the 1976 study of college English placement and the TSWE. Percentages are based on 770 students who were administered a 20-minute essay before beginning a first course in freshman English. The TSWE score ranges represent scores obtained at the time the students took the SAT when applying for college. Four institutions participated in the study. Two of the four institutions were located in metropolitan areas, one was located in a suburban community, and one was located in a small town. Because this sample of students may be systematically different from the remainder of students within the participating institutions, they are not necessarily representative of those freshman classes.

Table 15. Mean scores for adults tested in 1981-82 (page 26)

Data for this table are based on approximately 31 056 adults, 20 years of age or older, who took the SAT between October 1981 and June 1982.

Table 16. Predictive validity coefficients using SAT scores and high school record for prediction of freshman GPA, by sex and college curriculum (page 27)

The data on which these results are based were collected in the ATP Validity Study Service for colleges conducting studies relating SAT scores and high school record to freshman grade point average of their entering freshman classes. All studies of the whole freshman class, of male freshmen, of female freshmen, and of freshmen entering any of six selected college curriculums are included. (If a college conducted more than one study of any type, only the most recent one is included.) Although the validity coefficients reported are typical of those from past validity studies, the colleges contributing data may not be representative of the entire group of colleges that use the Admissions Testing Program.

Appendix B. Testing Terms

Appendix B contains brief definitions of testing terms used in the Guide. For a discussion of these statistical concepts on the level of a rigorous undergraduate course in tests and measurements, see *The College Board Technical Handbook for the Scholastic Aptitude Test and Achievement Tests.*

Correlation: the tendency for two measures or variables, such as height and weight, to vary together or be related for individuals in a group. If, as in the case of height and weight, people who are high on one variable (tall) tend to be high on the other (heavy), the correlation is said to be positive. As another example, months of practice and golf scores would have negative correlation, for ordinarily as the first variable is high (practice increases), the second tends to be low (score decreases). Correlation does not imply cause but only association.

Correlation coefficient: the customary index for expressing the degree of relationship observed between two sets of measures for the same group. The coefficient can range from −1.00, showing perfect negative correlation, through zero, indicating no correlation, to +1.00, showing perfect positive correlation. If, for example, the correlation coefficient between height and weight for a group of men were +1.00, knowing a man's height, one could predict his weight without error. But, if the correlation coefficient between height and weight were zero, one could not predict a man's weight, knowing his height, any more accurately than not knowing his height.

Most correlation coefficients of test scores and measures of academic success fall somewhere between zero and +1.00. Knowledge of an individual's score on one variable enables one to predict his or her standing on the other variable imperfectly but with greater accuracy than if the correlation were zero. The higher the coefficient, the less error likely in prediction.

Equating: a statistical procedure that puts the raw scores of newly introduced forms of a test on a continuing scale and compensates for variations in difficulty among various forms of the test. Equating often involves comparison of the performance of an old and a new group of candidates on the same test material. Sometimes a random sample of new candidates takes an entire old form of the test, or a representative sample of material from an old form may be added to a new form. By comparing the performance of past and current candidates on the old material, the difference in their ability levels can be observed and the mean and standard deviation of scores made on the new form can be adjusted to reflect this difference. Thus, in a single operation a formula is derived that permits scores made on the new form to be placed on the scale (see the entry for Scaling) and calibrates the scores on the new form so they are comparable to those from old forms, regardless of differences in difficulty. Because successive forms of the test are constructed with knowledge about the difficulty level for which equating must compensate, the adjustments are not very great.

Frequency distribution: a tabulation of scores from high to low, or low to high, showing the number of individuals who obtain each score or whose scores fall in each score interval. Frequency distributions are used to determine tables of percentile ranks.

Mean or arithmetic mean: the average.

Median: the score below which 50 percent of the cases in a score distribution fall. If the distribution of scores is distorted by the presence of a few aberrant cases of little importance, the median may be a better summary description of the group than the mean. If the distribution is symmetric, the median and mean will be almost identical. The median is also by definition the 50th percentile.

Norms: a statistical description of the performance on a test of a well defined group that serves as a reference with which to gauge the performance of other individuals who take the test. Most norms tables show in descending order various test scores and the percentage of people in the reference group who scored below each score level. Thus, knowing an individual's score, one can quickly determine how he or she compares with the reference group.

Obtained score: the score actually achieved by a person taking a test. It is considered to be the sum of the individual's true score plus the error introduced by the imperfect reliability of the test. This error can be either positive or negative so that an obtained score can be higher or lower than the true score. (See True score.)

Percentile rank: the percent of scores in a distribution that are lower than a particular obtained score. The remaining scores are at the same level or higher.

Raw score: the number of correct responses minus a fraction of the incorrect responses. The raw score is converted to a scaled score for reporting.

Reliability: the extent to which a test measures consistently, that is, the extent to which a person repeating the test or taking an alternate form would tend to get the same score, assuming that practice makes no difference. Reliability is usually expressed as a correlation coefficient and is comparable, roughly, to the correlation of a test with a perfectly parallel form of the same test.

The concept of reliability can be illustrated as follows. Imagine two yardsticks, one made of wood and the other of a tape that is subject to stretching. Obviously, the wooden yardstick will be more reliable because it will give more consistent results if an object is measured with it repeatedly. Some error is introduced into all measurements, whether tests or machine shop measurements, by the imperfect reliability of the measuring instrument.

Rescaling: Realigning the system for transforming raw scores to reported scores for a test or testing program. Achievement Tests were rescaled for the following years: 1965 through 1972, 1974, 1976, and 1978 through 1980.

Scaling: a means of defining a system for transforming raw scores to reported scores for a test or a testing program.

Standard deviation: a measure of the spread or extent of variability of a set of scores around their mean. The standard deviation reflects the degree of homogeneity of the group with respect to the variable in question. That is, the less the dispersion of scores, the smaller will be the standard deviation.

Standard error of the difference: an indication of the extent to which the difference between the scores of two people on the same test or the scores of one person on two different tests may represent error due to the unreliability of the test. The user can be reasonably confident that the higher score represents greater ability or achievement as measured by the test if the difference between two scores exceeds 1.5 times the standard error of the difference for the test.

Standard error of measurement: an index of the extent to which students' obtained scores differ from their true scores. It is expressed in score units of the test. Intervals extending one standard error above and below the true score will include 68 percent of candidates' obtained scores. Similarly, intervals extending two standard errors above and below the true score will include 95 percent of the candidates' obtained scores.

True score: a hypothetical concept indicating what an individual's score on a test would be if there were no error introduced by the measuring process. It is thought of as the hypothetical average of an infinite number of observed scores with the effect of practice removed.

Validity: an indication of the extent to which a test or other measure does the job for which it was intended. There are several kinds of validity. Predictive validity is the extent to which test scores, for instance, are able to predict a criterion variable such as grades or faculty ratings. Validity is expressed as a correlation coefficient between the predictor variable, such as the SAT scores, and the criterion variable. Validity coefficients, like all correlation coefficients, are heavily influenced by the extent to which the individuals studied are spread out on the predictor measure and on the criterion measure. If the range of ATP scores or freshman year grades is restricted, the correlation between the two will be smaller. In practice, the range of scores for admitted students is almost always smaller than that for the total applicant group. Therefore, a validity coefficient based on admitted students will underestimate the usefulness of the test scores as an aid in selecting among applicants.

Variables other than test scores, including especially the high school average, also have predictive validity. The highest predictive validity is generally obtained by combining test scores and high school class rank or grade average, weighted in accordance with the results of a multiple regression analysis.

Appendix C. Guidelines on the Uses of College Board Test Scores and Related Data

The College Board has prepared and distributed widely a statement entitled *Guidelines on the Uses of College Board Test Scores and Related Data*. These guidelines are addressed to all those who use College Board tests and related data or are otherwise concerned about their use. They describe how the College Board as sponsor of test services interprets its own responsibilities in relation to the public and to members of users and what additional responsibilities it believes the users of these services have. The guidelines also state the conditions the Board regards as appropriate for the several uses of its tests, call attention to certain practices it regards as inappropriate, and set forth the procedure to follow in questioning the use of test scores and related data. Portions of this statement are printed below. To request a free copy of the complete *Guidelines*, write to College Board Publications, Box 886, New York, NY 10101.

The College Board should:

Adhere to the highest standards in the development and administration of its tests and related services, giving careful attention to such generally accepted standards as those embodied in *Standards for Educational and Psychological Tests and Manuals* (1974), promulgated by the American Psychological Association in association with the American Educational Research Association and the National Council on Measurement in Education.

Provide those who use its testing services—counselors, admissions officers, school and college faculty members and administrators, and test takers themselves—with full information about the purposes and nature of the services.

Advise educational institutions and agencies about what its tests and related services are designed and able to do for them and their students or clients and, correspondingly, what responsibilities and capabilities they should be prepared to exercise if they use such services.

Assure appropriate use of its tests and related services by maintaining test instruments that are current and relevant to the domains they measure, publishing information essential to understanding and using the services properly, communicating and consulting on a regular basis with using institutions and agencies, and following up instances thought to represent misuse with advice and assistance to the users in question.

Assure the appropriateness and fairness of its tests through the engagement of faculty members at both secondary and postsecondary levels as appropriate, in the construction and review of the tests and through periodical surveys of curriculum content in educational institutions.

Achieve fairness and sensitivity to the concerns of minorities, women, and other subgroups of the test-taking population through special reviews of test questions.

Maintain for each of its test programs procedures for seeking advice and criticism from program users, students as well as institutions, about the quality and adequacy of the services provided.

Maintain effective procedures for protecting the privacy of individual test takers, releasing information that serves to identify them only with their consent.

Respect the interests of educational institutions, state departments of education, and consenting groups of institutions or agencies, releasing identifiable summary or aggregated data pertaining to them only to people explicitly authorized to receive such information.

Maintain effective procedures for verifying the scores of test candidates who question their accuracy and for responding with care to candidate queries or complaints about particular test questions or test administration procedures.

Schools, colleges, and scholarship agencies that use College Board test scores and other related information should:

Assign responsibilities involving test use to people knowledgeable about educational measurement, including current literature regarding the purposes, content, statistical characteristics, capabilities, and limitations of any test in use or under consideration.

Provide those who may have occasion to take tests with full information about them, including which ones are required or available and how the information they yield will be used.

Protect the privacy of test candidates by treating confidentially, in accordance with the Family Educational Rights and Privacy Act, scores and other information derived from tests they take.

Make use of College Board scores and related data with discretion and only for purposes they are capable of serving.

When College Board tests are used for counseling purposes, counselors should:

Advise counselees on what tests they may need to take in pursuing their educational objectives, when and where they might conveniently take the tests in view of institutional requirements, testing schedules, and their own personal schedules, and how they can most constructively interpret their scores in their own best interests.

Explain the limitations as well as the capabilities of tests: that tests, like all measuring devices, are not perfectly precise and should not be treated as though they are; that admissions test scores are useful as one means of predicting academic performance in college but are not infallible predictors and should be considered along with other relevant information.

Inform students that admissions test scores are intended to be used, and are used by most colleges, in conjunction with secondary school records and other relevant information, with the scores providing a useful uniform measure for all students, in contrast to school records, which are based on widely varying grade standards.

Release the scores and other information derived from a test a counselee takes only with the counselee's consent, if the released information could serve to identify him or her.

When institutions use College Board tests and related data in conjunction with other information for recruiting purposes, as in the case of the Student Search Service, they should:

Seek to recruit only those students they are capable of serving and

Identify the source of the information (e.g., the College Board's Student Search Service) at the time they first communicate with prospective applicants.

Use the information only for their own recruiting purposes, consistent with assurances given to test candidates by the College Board.

Provide prospective applicants with relevant information about the institution, including its environment, students, and programs, the opportunities it provides for financial assistance and for placement and/or credit by examination, and the qualifications required for special academic programs.

Provide prospective applicants with relevant and helpful information about the characteristics of enrolled students and recent graduates.

Publicize admissions requirements, procedures, and deadlines as they relate to College Board services and ensure that such materials are readily available to prospective applicants.

When colleges use College Board tests for selection purposes, the responsible officials or committee members should:

Know enough about tests and test data to understand their proper use and their limitations.

Consider test scores and related data from the College Board's Admissions Testing Program as supplemental to the secondary school record and other information about applicants in assessing their ability to undertake college-level studies, recognizing that a combination of predictors is almost always better than a single predictor.

Validate data used in the selection process regularly (e.g., every three years) to ensure their continuing relevance, using if desired the Validity Study Service of the College Board, which is available without charge.

Take into appropriate consideration predictions of performance for applicant subgroups (men and women, ethnic groups, foreign students, curricular groups, adults) in developing equitable admissions policies and practices.

View admissions test scores as current and approximate indicators rather than as fixed and exact measures of a student's readiness for college-level work.

Maintain adequate procedures for protecting the confidentiality of test scores and other admissions data.

When systems or groups of colleges use College Board tests for selection (admissions) purposes, the officials responsible for the group or system should

Know enough about tests and test data to understand their proper uses and their limitations.

Collect and consider recent admissions validity data for each individual college in the group or system and conduct appropriate validity studies for the system or group as a whole or for major subgroups.

Consider test scores in conjunction with information about the secondary school record and other information about applicants in assessing their ability to undertake college-level studies, recognizing that a combination of predictors is almost always better than a single predictor.

Conduct appropriate studies to ensure that uniform standards can apply and are appropriate to the populations of students served and to the missions of the colleges.

Take into reasonable consideration predictions of performance for applicant subgroups in developing equitable admissions policies.

Require that individual institutions validate data used in the admissions process and conduct appropriate system or group studies regularly (e.g., every three years) in order to ensure the continuing relevance and appropriateness of the information used in the combinations established for the admissions policies.

Before determining the admissions policies to be adopted for the group or system of colleges, allow sufficient time and opportunity for representatives of the individual institutions to consider and discuss possible policies and to suggest alternative policies, especially as these relate to their institutions.

When introducing or revising admissions policies, allow sufficient lead time and provide considerable notice to schools and students so they can take the new policies into account when planning school programs and curricular offerings.

Avoiding the Misuse of Test Scores

Abiding by the foregoing guidelines will help ensure that scores and related data are used appropriately and defensibly from an ethical and technical standpoint. Since the decisions and judgments that test scores influence may have significant personal and social consequences, however, every care should be taken to avoid practices involving the use of scores that might turn tests from aids to informed judgments and decisions into instruments of injury and injustice. When tests are used properly, the result should be decisions that are better in significant respects than they would have been without the influence of the scores.

The Council on Entrance Services of the College Board has found that while a few specific uses of tests meet with either universal approval or disapproval, most, on the contrary, are the focus of at least some debate about their merits or flaws. The disagreements sometimes turn on differences in the specific circumstances that may accompany a particular practice. Judgments about the use of test scores to screen applicants, for example, may be influenced by such considerations as how valid they are as predictors, how feasible it is to use other information instead of or in addition to the scores, how many applicants can be accepted by the school, and whether qualified students might be screened out by a minimum score level. Even when all the details of a given practice are held constant, however, there are sometimes honest differences of opinion about the essential rightness or wrongness of the practice. The technical issues raised are ultimately likely to be of less significance than questions of public policy and institutional prerogatives.

The council has concluded that the most helpful advice it can offer users as a guide to avoid unjust or otherwise injurious practices is to consider carefully the limitations of any test on those tested and to guard specifically against using tests to serve purposes they are neither intended to serve nor capable of serving. The following are examples of uses that should be avoided:

1. Using the SAT as a measure of the overall performance of either teachers or of schools.
2. Encouraging the belief that the SAT or other College Board tests measure a person's worth as a human being.
3. Using test scores as the sole basis for important decisions affecting the lives of individuals, when other information of equal or greater relevance and the resources for using such information are available.
4. Using SAT or other College Board test scores in ways that are not subject to appropriate consideration of their validity.
5. Providing inadequate or misleading information about the actual influence of test scores on particular judgments or decisions.
6. Requiring or recommending that certain tests be taken when the scores are actually not used at all or are used to a negligible extent.
7. Interpreting the scores on any test without regard to the concepts or skills it measures.

Writing the College Board

Comments, complaints, inquiries, and suggestions about the use of College Board test scores and related information are welcome and should be addressed to the appropriate College Board Regional Office. (See the back cover for a list of the College Board Regional Offices.)

Copyright © 1977, 1981, 1988 by College Entrance Examination Board. All rights reserved.

Test Date Formula

Please see the current ATP *Registration Bulletin* for test dates and registration deadlines. Deadlines for countries other than the United States are given only in the *International Edition* of the *Registration Bulletin*. (See the *New York State Edition* for test dates in that state.) The test dates are determined by the formulas given below.

[redacted]

College Board Regional Offices

New England
Middle States
West (also includes Alaska and Hawaii)
Midwest
Alaska
Texas
Hawaii
Southwest
South
Puerto Rico

Middle States: Suite 410, 3440 Market Street, Philadelphia, PA 19104-3301 (215) 387-7600
Midwest: Suite 605, 500 Davis Street, Evanston, IL 60201-4697 (312) 866-1700
New England: 470 Totten Pond Road, Waltham, MA 02154-1982 (617) 890-9150
South: Suite 200, 17 Executive Park Drive, N.E., Atlanta, GA 30329 (404) 636-9465
Southwest: Suite 922, 211 East Seventh Street, Austin, TX 78701 (512) 472-0231
West: Suite 480, 2099 Gateway Place, San Jose, CA 95110 (408) 288-6800
Suite 705, 4155 East Jewell Avenue, Denver, CO 80222 (303) 759-1800

In Puerto Rico, inquiries should be directed to
The College Board, Banco Popular, Suite 701, Hato Rey, Puerto Rico 00918 (809) 759-8625
Mailing Address: Call Box 71101, San Juan, Puerto Rico 00936-1101

In Alaska and Hawaii, inquiries should be directed to the Western office at the California address.

1:0

Bias in Testing

NANCY S. COLE *University of Pittsburgh*

ABSTRACT: *The problem of test bias has recently received tremendous scientific and public scrutiny. Cole reviews the approaches that have been undertaken to detect cultural, content, prediction, and selection bias in mental tests. This includes analyses of subtle differences in the content of test items to which individuals react differently and the implications of statistical differences in predictions from test scores. She argues that questions of bias are fundamentally questions of validity. A distinction is made between validity on one hand, and the question of whether a test should be used, even if valid, on the other. The author concludes that although on the technical side many things have been learned about the details of test bias, such research has not provided answers to social policy questions that must be decided regardless of whether tests are involved.* —The Editors

Some of the most prominent issues associated with testing in recent years have involved questions of test bias. The possibility of bias in tests has been a major focus of test critics, the courts, test developers, and scholars of testing alike. The focus has resulted in much public debate and scholarly writing, but a large gap continues to exist between the public concerns and the concerns of technical scholars of testing.

The issue of test bias has gained importance as much because of the social policy issues with which it has become intertwined as because of its own intrinsic importance. Shepard (Note 1) aptly noted one prominent issue associated with test bias:

> One reason that bias in mental testing is so volatile an issue is that it involves the specter of biological determinism, i.e., whether there is a large difference in intelligence (IQ) between black and white Americans which can be attributed largely to inherited differences. (p. 3)

The reasons for observed group differences in test scores and the possible social policy implications of different reasons are but one set of the volatile issues with which the issue of test bias has become at least marginally related. Others include issues such as the provision of educational and job opportunities for minorities (Novick, this issue; Tenopyr, this issue), the appropriateness of particular types of educational intervention (e.g., assigning children to EMR classes—Bersoff, this issue), and the use of tests in the certification of a high school diploma (Lerner, this issue). Thus, while scholars study the issue of bias largely for its own sake, the public continues to grapple with the much harder and broader issues of social, economic, and educational policy within which the issue of bias is often embedded.

This article seeks to provide an overview of the scholarly research on test bias—how the questions have been approached and what results have been found. A further purpose is to distinguish the questions this research can and cannot answer. For example, the large amount of technical work on bias done in the last 15 years has produced noticeable effects on the test construction and data gathering procedures used with many major, widely used tests (especially those with high volume and frequent revision schedules such as standardized achievement tests and college admissions tests) and probably has resulted in some improvements in test quality. However, this work cannot be said to have significantly clarified the public controversies or to have resolved many issues associated with bias in the public mind.

The Validity Basis of Technical Bias Approaches

VALIDITY AND ITS PARTS

The guiding tenet for the technical scholar of testing has been the validity tenet. In the definitive treatment of test validation, Cronbach (1971) wrote that

> narrowly considered, validation is the process of examining the accuracy of a specific prediction or inference made from a test score. . . . More broadly, validation examines the soundness of all the interpretations of a test. (p. 443)

Requests for reprints should be sent to Nancy S. Cole, Educational Research Methodology, University of Pittsburgh, Forbes Quad 5C01, Pittsburgh, Pennsylvania 15260.

Vol. 36, No. 10, 1067–1077

Copyright 1981 by the American Psychological Association, Inc.
0003-066X/81/3610-1067$00.75

Thus, the validation process involves collecting information concerning the accuracy of interpretations of test scores. When group differences exist on these test scores, the validity-based bias question concerns whether such differences in the scores accurately reflect group differences in the characteristic the scores are supposed to represent (ability levels, preparation for college, job capabilities, etc.). It is therefore not surprising that when group differences in test scores were put forward by some as evidence of bias in tests, testing scholars responded with studies to discover if such test score differences were accurate (valid) or not accurate (biased).

Test validation has traditionally been segmented into three types, each aimed at a different type of use or interpretation of a test score. When a test score is used to predict a criterion performance level, then the validation concern is the accuracy of that prediction, and the label criterion-related validity or predictive validity is applied. The study of bias in selection grew directly out of predictive-validity approaches to validate predictions of performance in educational or job settings for various groups of concern. When a test score is to be used to represent a construct or hypothetical characteristic of an individual, the concern is with how accurately the score represents the construct, and the term construct validity is used. Various construct validity approaches have been used to study bias in IQ tests, for example, and all statistical item bias procedures represent construct validation approaches. When a test score is used to reflect performance in some well-specified content domain, then the adequacy of the test as a sample from that domain is at issue, and the process of validation is labeled content validation. Judgmental processes associated with content validation have also been prominent in bias considerations.

More recent work on validity has attempted to counter this three-part approach to validity by bringing all three parts under the construct validity umbrella (Cronbach, 1980; Messick, 1975). These authors question whether the meaning given a test score is ever so narrowly confined to a prediction or a content domain as to preclude the need for a broader construct-type understanding of the meaning of the score. From this perspective, the different methods traditionally associated with the three types of validity (and different aspects of bias) can be viewed more appropriately as different types of information relevant to a better understanding of the meaning (or bias) of a test score. The more unified conception of validity makes clear that, in understanding the possibility of bias in a test score, no one bias approach should stand alone (Shepard, Note 1). Various types of information about bias—whether from predictive validity perspectives, methods traditionally associated with construct validation, or judgmental procedures—provide different types of evidence. Combining these types of information gives a more complete understanding of possible bias than using any one type alone.

VALIDITY AND ITS LIMITS

As already noted, technical research on bias has not answered many public concerns related to bias. The reason for this gap can be seen in recent considerations of the limits of validation theory. Messick (1975) distinguished two questions:

> First, is the test any good as a measure of the characteristic it is interpreted to assess? Second, should the test be used for the proposed purpose? The first question is a technical and scientific one and may be answered by appraising evidence bearing on the test's psychometric properties, especially construct validity. The second question is an ethical one, and its answer requires an evaluation of the potential consequences of testing in terms of social values. (p. 962)

Cronbach (1980) made the same distinction between finding out what a test measures and making a policy decision involving the question of whether it should be used. In an example noting the additional policy issues involved to address selective educational admission policies, he noted that "the whole selection system is to be justified, not the test alone" (Cronbach, 1980, p. 102).

As members of a scientific community, testing scholars clearly value scientific evidence on whether a test is accurate for some intended inference. For example, an intelligence test might accurately (validly) identify mentally retarded children. However, if a test is accurate (valid), validity theory does not say whether the use of the test, or the whole system in which the test use is embedded, will produce a social good or a social evil. Thus, the use of a valid intelligence test to identify mentally retarded children for assignment to a special educational intervention does not ensure that the special intervention will provide the education the children need. Even if the test is valid for the inference about mental retardation, the actions taken on the basis of the test scores may be either positive or negative ones for even the correctly identified child.

The distinction between the validity questions concerning bias and questions of an educational or social policy are especially important because scholars studying validity questions of bias (and the courts hearing their testimony) have sometimes assumed that the answers to the validity questions would also provide answers to questions of social policy. A valid test may be used to produce results that are educationally or socially negative, just as an invalid test may be used in a way that produces positive results. Testing scholars and the courts must clearly recognize that test validity and the appropriateness of social or educational policies are separate issues. Thus, even as the scholarly community strives to assume with (and belief in the value of) validity-type evidence about test bias, it must not be blinded to the limitations of this evidence in answering the essentially different questions of the relative desirability of alternative social policies. The scholar of test bias has dealt only at the level of test validity, but Messick's second question, "Should the test be used?" or Cronbach's, "Should the policy be implemented?" remain the vexing issues facing the courts and the public.

Bias in Selection

Validation in selection situations has traditionally been viewed from the perspective of predictive validity. A predictor such as a test score that accurately predicts eventual criterion performance (such as college grades or job ratings) is viewed as valid for selecting individuals who can eventually perform adequately on the criterion. A test has predictive validity to the extent that those scoring high on the test do well on the criterion (in college, on the job), and vice versa. When there are differences in predictor (test) scores between groups such as blacks and whites or men and women, the prediction bias question involves whether these group differences are accurately reflected in criterion differences or whether they represent bias.

DIFFERENTIAL PREDICTIVE VALIDITY

The predictive relationship of test scores to criterion scores is often expressed in terms of a regression line relating the test score to the criterion. When the regressions computed separately for special groups of interest coincide, the test is deemed to give a fair[2] prediction for each group; when the regressions diverge, scores do not give the same predictions for each group and may be considered biased (Anastasi, 1968; Bartlett & O'Leary, 1969; Cleary, 1968; Cohen, 1968). Thus, group differences in predictive validity or differential predictive validity have become labels to describe a type of bias.

Empirical studies of prediction differences between groups have investigated differences in the level of predictor-criterion correlations and in the regression equations within different groups (American College Testing Program, 1973; Campbell, Crooks, Mahoney, & Rock, 1973; Jensen, 1980; Kallingal, 1971; Kirkpatrick, Ewen, Barrett, & Katzell, 1968; Linn, 1978; Stanley, 1971; Stanley & Porter, 1967; Temp, 1971). Complications such as those presented by differential reliability in groups have been noted (Linn & Werts, 1971), and various modifications of the regression approach have been proposed (Einhorn & Bass, 1971; Jensen, 1980; McNemar, 1975).

The empirical results of differential predictive validity studies have been complicated by the variety of ways group differences can occur. There can be differences between groups in predictor-criterion correlations or in predictor reliabilities. However, empirical studies of group differences in reliability (American College Testing Program, 1973; Green, 1978; Jensen, 1980) have reported only small reliability differences between groups. The regressions may differ in slope, intercept, or standard error of estimate, and statistical differences in some feature of the regression have frequently been reported. However, a consistent finding has been that the use of a single regression equation based on combined data from black and white groups results in the overprediction of the performance of black students. The preponderance of the results led Cleary, Humphreys, Kendrick, and Wesman (1975) to conclude that evidence of differential validity negative to minority groups had not been established in educational testing. Schmidt, Berner, and Hunter (1973) surveyed similar studies in the employment literature and found no bigger than chance differences in predictor-criterion correlations between black and white groups. Thus, from a large number of educational and employment studies, the most common conclusion has been that many tests predict various educational and employment performances about as well for minority groups (blacks and women being by far the most frequently studied minority groups) as for majority groups.

---

[2] In this article the words *fairness* and *bias* are used as simple converses of each other. Some authors have assigned different meanings to the two words, with bias used for the technical (validity) issue and fairness for social policy issue. The distinction is handled here by distinguishing validity issues from the social policy issues but keeping the familiar meanings of the two words.

PREDICTION VERSUS SELECTION POLICIES

The approaches to prediction bias described in the preceding section are limited to the validity perspective and do not address Messick's ethical question about what should be done in selection. In Cronbach's terms, the predictive validity approaches do not justify the whole selection process. The importance of recognizing this distinction between validity and appropriate selection policies can be seen in discussions about whether any variable contributing to predictive accuracy should necessarily be used in selection. For example, in some instances, racial-ethnic identity may be a valid (accurate) predictor of some performance. If one goes no further than validity concerns, the conclusion could be reached that any valid predictor should be used, as Jensen (1980) has argued. However, when the myriad of values and judgments involved in determining an appropriate selection policy are recognized, it becomes clear that the issue of appropriately using a variable (such as racial-ethnic identity) in selection primarily involves value judgments, not technical validity judgments. Again the distinction between the validity of a variable as a predictor and appropriate social policies (in this case, selection policies) must be clearly recognized.

Questions about the role of value judgments in determining appropriate selection policies and the distinction between test validity and appropriate selection policies have arisen from another type of methodological study as well. Thorndike (1971) notes that it is possible under the regression approach to prescribe the selection of smaller proportions of one group than another even though the potential success rates in the two groups (if all applicants had been admitted) might be more similar. He then suggests the possible policy of selecting from different groups in proportion to past success rates in the groups. Thorndike's rule represents a reasonable possibility for a fair selection policy, but it differs from the selection rule derived from the predictive validity standpoint alone. Darlington (1971) notes various possibly reasonable (but different) notions of bias and suggests a procedure to build favoritism into a selection system in which one group is favored based on some value judgment about fairness. Cole (1973) proposes a rule for fairness in selection that involves selecting equal proportions of those who would eventually qualify on the criterion in each group (requiring the conditional probability of selection given success on the criterion to be equal across groups). Several authors have analyzed these approaches and compared them to the regression approach (Jensen, 1980; Linn, 1973; Petersen & Novick, 1976; Schmidt & Hunter, 1974). These different approaches make it clear that prediction fairness (the validity issue) is but one of many possible conceptions of selection fairness as a social policy (Messick's second question). Further, different social values about what constitutes fairness can lead to different technical conclusions about whether a test used in a selection situation is or is not fair.

Gross and Su (1975) and Petersen and Novick (1976) point the way to statistical decision-theoretic formulations of selection in which social values are made explicit as "utilities." By value or utility is meant the quantified, relative importance one attaches to specific selection outcomes. For example, if one considered it socially important to have black police officers on the police force, one might "weight" this aspect of employee selection relative to actual job performance. Under these statistical models, it is possible to compute the selection rule that gives the optimal selection outcome according to a specified set of values. Different people with different values would have different optimal selection rules. These decision-theoretic models make explicit some of the value issues raised by Thorndike, Darlington, and Cole, but limit the type of values that could be expressed. Sawyer, Cole, and Cole (1976) extend this approach to allow broader classes of values to be expressed, particularly values for group outcomes, such as those concerning fair treatment to groups of potentially successful applicants that may not extend to a comparable concern for those potentially unsuccessful. Under this broader framework, both the Thorndike and the Cole proposals could be seen as logically consistent decision-theoretic solutions under particular types of values.

In a thoughtful note culminating the decision-theoretic approach to selection bias, Cronbach (1976) warns that the issues of bias have not been solved and "will not be settled by mathematical specialists" (p. 31), but by carefully examining and debating the different value positions expressed. Thus, the statistical approach helps to distinguish

[illegible]

[illegible]

Bias in Internal Test Structure

[illegible]

[illegible]

[illegible]

[illegible]

Statistical Item Bias Methods

[illegible]

[illegible]

is that the rest of the items measure, the biased item behaves differently. (pp. 17–18)

In this section, statistical item bias methods are divided into two major categories—those based on Item × Group interactions and those using more sophisticated item response theory approaches. The discussion of these two approaches to item bias draws heavily on Cole and Nitko (1981). In a recent review of item bias techniques, Rudner, Getson, and Knight (1980) provide a more detailed review of the various techniques and give more attention to the technical formulations than space allows here.

ITEM × GROUP INTERACTION APPROACHES

Early item bias approaches applied analysis of variance techniques to item data for several groups to examine Item × Group interactions (Cardall & Coffman, 1964; Cleary & Hilton, 1968). Prominent Item × Group interactions would mean that the items were operating in different ways in different groups and hence not measuring the same thing for different groups in the construct validity sense. The same type of effect could be seen by correlating item difficulties (or some transformation of item difficulties) for different groups. The analysis of variance and correlational approaches produced summaries of the overall degree of anomalous item behavior across groups, but further methods were needed to explore which particular items were the most anomalous ones. These item-level approaches were provided by Angoff and Ford (1973) and Angoff and Sharon (1974), who plotted the item difficulties (actually an inverse normal transformation of the item difficulties) for one group against the other. In these plots, consistent group differences (not necessarily labeled bias) were reflected in deviation of the entire cluster of item points from the diagonal; potentially biased items would then be the outliers from the cluster of items and indicate items that favored one group more than the bulk of items on the test. Fishbein (Note 2) used a statistical test to look for item difficulty differences greater than the average difficulty difference on the whole test, and Veale and Foreman (1975) proposed comparing response rates to incorrect answers, looking for differences in response patterns.

ITEM-RESPONSE THEORY APPROACHES

Item-response theory supplies a statistical model for describing the relationship of the ability level of the examinee to the probability of a correct response to an item. Using such a model, items can be characterized (described) by the form of the curve relating ability level to probability of correct response. Such "item characteristic curves" can be computed for different groups and compared. As with the factor structure and Item × Group interaction approaches, group differences in item-characteristic curves are not expected if the test measures the same construct in the different groups. Consequently, differences in item-characteristic curves in different groups would be evidence against the construct validity of the test across groups and hence, in this sense, evidence of bias. Lord (1977) notes ways in which comparing item-characteristic curves across groups is more appropriate than the Item × Group interaction procedure.

Green and Draper (Note 3) examine the percentage of correct responses for each total score level (similar to empirical item response curves) in samples of minority and majority students. Lord (1977, 1980) and Rudner (Note 4) apply the most complex of current item-response theory models (a three-parameter logistic model) to estimate and compare the item characteristic curves in different groups. Both these procedures require large sample sizes (probably over 1,000 per group) and are therefore practical only for large-scale testing programs and even then only after a test has been constructed and is in operation. Durovic (Note 5) and Wright, Mead, and Draba (Note 6) propose use of the simpler item-response model (the one-parameter Rasch model) requiring smaller samples to examine bias. Scheuneman (1979, Note 7) proposes a statistical test for group differences based on a kind of grouped empirical item-response curve that is a smaller sample alternative to the more elegant but complex item-response theory models. (See Ironson, Note 8, for a detailed discussion of possible variations and modifications of this statistic.) Of the various methods, the complex three-parameter logistic model procedure used by Lord and others seems preferable because of its theoretical completeness and its performance in empirical studies as well (Ironson, Note 9; Subkoviak, Mack, & Ironson, Note 10). The other methods are of considerable interest, however, as small-sample options, especially in the early stages of the test construction process, before data on large samples become available.

RESULTS FROM THE STATISTICAL STUDIES

Most studies using one or more of the item bias procedures have reported at least some, and oc-

[illegible]

Judgmental Analyses of Item Bias

[illegible]

FACIAL BIAS

In understanding the confusing results of judgmental analyses of item bias, it is useful to distin-

[illegible]

CONSTRUCT-VALIDITY-BASED JUDGMENTS

[illegible]

dices. It seems important, then, to distinguish the types of bias concerns that may be facial concerns alone from those that have construct validity implications.

ITEM BIAS JUDGMENTS IN TEST CONSTRUCTION

Statistical item analyses of item difficulty, discrimination, and distractors are a standard part of the test construction process as indicators to alert test constructors to potential item flaws. Editorial judgment is then used to consider and possibly confirm the nature of the flaw and to correct it. Statistical item bias indicators and item bias judgments can be seen as extensions of traditional item analysis. The complications, however, of the addition of these bias concerns to test development are that (a) the statistical procedures require substantial numbers in each special group of concern, considerably inflating the numbers typically needed for item tryouts, (b) judgments of facial bias would ideally be done prior to tryout, and (c) judges of construct validity–based bias would need dual expertise in the construct and in the groups of concern, which would probably require an addition of new people with special expertise to the test development process. Even with these complications, some test publishers have implemented statistical and judgmental bias reviews as part of the test development process.

Concluding Remarks

Scholarly inquiries concerning bias in testing over the past decade have resulted in a massive technical literature that is necessarily only sketchily surveyed here. Although much has been learned, the failure to produce clear and unambiguous results of the type originally envisioned has left many people somewhat dissatisfied. Perhaps it is useful to review the expectations, the findings, and the social context of test bias research to better understand what has and has not been learned, as well as what has been unsatisfying to many scholars.

The concern with test bias has arisen out of a broad social concern with equitable treatment of special groups in this society. This concern has been embedded in a number of specific questions of social policy involving a variety of other complex issues in addition to the issue of equity to special groups. For example, the nation has faced several difficult social policy issues:

1. In selecting for employment, what is the best policy to combine concerns for compensating for past wrongs with current employer needs and current individual rights?
2. What role should selective admissions play in higher education and how should it be balanced with the desire to broaden opportunities?
3. What form should the education of children with handicaps, such as mental retardation, take?
4. How should we deal educationally with people who are behind desired levels of learning?
5. How should we deal socially and educationally with people for whom English is not a first language?
6. What should a high school diploma mean and who should have control of certifying its meaning?

Almost hidden inside these complex issues is a test, and sometimes the fairness of that test has surfaced as an apparently crucial issue in the social policy question. Thus, many people motivated by [illegible] critical questions of social policy began to pursue scholarly research on test bias at the technical validity level. It is interesting that these policy-level motivations were generally of two opposing types. On the one hand, much test bias research has been motivated by the belief that tests are socially constructive and by the desire to defend them from attack. On the other hand, perhaps as much research has been motivated by a concern that tests might be contributing, through bias, to a harmful situation and the desire to find any bias that exists and to eliminate it. Thus, many researchers in the test bias area can be roughly characterized as belonging to one of two groups: the defenders or the reformers. The defenders expected to find no bias and to contribute to the solution of several social issues by eliminating the specter of bias from the public debate. The reformers expected to find bias and to contribute to the solution of the same issues by forcing reforms to correct bias. However, neither group has found perfect satisfaction in its quest.

From the defenders' point of view, although no clear picture of substantial bias emerged, enough hints of possible trouble were uncovered to suggest that some subtle issues of bias might need attention. Thus, although predictive studies showed strong evidence of overprediction of some minority groups' performance (the opposite of the type of bias expected by the reformers) and the structural relationships of scores were typically quite similar, frequent although inconsistent differences in the regression-prediction systems were found, and items showing statistical anomalies between groups were identified, if not always understood. The defenders could then ignore the inconsistencies and conclude, as Jensen (1980) did, that there is no evidence of bias in mental tests or continue to study the sticky complications and feel unsatisfied.

The reformers found equal dissatisfaction. The bias uncovered by research was too small and too subtle to make clear what action, if any, was needed for reform, yet there were enough provocative findings to convince the reformers that all was not totally well. Probably the most important area needing reform (and reforms were therefore implemented) was not even in the area of technical validity, but in the area I have labeled facial bias. Concerns with the appearance of bias and with potentially offensive content have been effectively raised and have resulted in clear reforms at least in some parts of the testing industry. However, such facial bias has apparently had little effect on any group's test scores, so reforming the tests has not eliminated the concomitant social issues.

Both defenders and reformers originally seem to have expected their technical discoveries to have implications for social policy questions of the type raised above. However, today the bias area seems unsatisfying to both groups both because they did not find the clear results they expected and because what was learned did not provide answers to the social policy questions that motivated the research. It now appears that there will be no "bias bombshells" with a clear, decisive impact on the issues of social policy with which many test uses are associated.

Although the bias scholars seem to have failed as influencers of social policy, they have not necessarily failed as researchers at the technical level. On the technical side, many important things have been learned, and there is the promise of more to come from the areas begun under the label of test bias research. That promise seems strongest (a) in understanding the subtle differences in the content of a stimulus (such as a test item) to which individuals react differently, and (b) in understanding the conceptual and psychological implications of statistical differences in items (such as item characteristic curves or predictions). We have begun to learn about small, not large, effects that are subtle, not obvious—effects that will likely have implications, when we finally understand them, for the education and testing of many individuals whether they are members of a minority group or not. In fact, it seems likely that investigators continuing these lines of research will come in the future to think less in terms of ethnically or sexually identified groups and more in terms of individual differences in test taking.

So when the educator or psychologist is asked what has been learned from test bias research, the answer must come in several parts. First, we have learned that there is not large-scale, consistent bias against minority groups in the technical validity sense in the major, widely used and widely studied tests. Second, we have learned that the lack of such bias means neither that the use made of the tests is necessarily socially good nor that improvements in the tests cannot be made. Third, we have learned that there are still many subtle aspects of the testing situation that we do not adequately understand and that are promising areas for future research to increase that understanding. However, these areas are not likely to yield results with a direct impact on sociopolitical policy decisions. Finally, and actually foremost, we have learned that whether or not tests are biased, their role is only a small part of the complex social policy issues facing the legislatures, the courts, and the citizenry at large. To pretend that these broader issues are essentially issues of test bias is to be deceived. These policy issues require decisions about values that must be made whether or not tests are involved.

REFERENCE NOTES

1. Shepard, L. A. Definitions of bias. In *Test item bias methodology: The state of the art.* The Johns Hopkins University National Symposium on Educational Research, Washington, D.C., November 1980.
2. Fishbein, R. L. *An investigation of the fairness of the items of a test battery.* Paper presented at the meeting of the National Council on Measurement in Education, Washington, D.C., March 1975.
3. Green, D. R., & Draper, J. F. *Exploratory studies of bias in achievement tests.* Paper presented at the meeting of the American Psychological Association, Honolulu, September 1972.
4. Rudner, L. M. *An approach to biased item identification using latent trait measurement theory.* Paper presented at the meeting of the American Educational Research Association, New York, 1977.
5. Durovic, J. J. *Test bias: An objective definition for test items.* Paper presented at the meeting of the Northeastern Educational Research Association, Ellenville, N.Y., October 1975.
6. Wright, B. D., Mead, R. J., & Draba, R. E. *Detecting and correcting test item bias with a logistic response model* (Research Memorandum No. 22). University of Chicago, Statistical Laboratory, Department of Education, 1976.
7. Scheuneman, J. *A new method of assessing bias in test items.* Paper presented at the meeting of the American Educational Research Association, Washington, D.C., April 1975.
8. Ironson, G. H. Chi square and latent trait approaches to the measurement of item bias. In *Test item bias methodology: The state of the art.* The Johns Hopkins University National Symposium on Educational Research, Washington, D.C., November 1980.

9. Ironson, G. H. *The comparative validity of classical and latent trait approaches to the measurement of item bias.* Paper presented at the meeting of the American Educational Research Association, Los Angeles, April 1981.
10. Subkoviak, M. J., Mack, J. S., & Ironson, G. H. *Item bias detection procedures: Empirical validation.* Paper presented at the meeting of the American Educational Research Association, Los Angeles, April 1981.
11. Strassberg-Rosenberg, B., & Donlon, T. F. *Content influences on sex differences in performance on aptitude tests.* Paper presented at the meeting of the National Council on Measurement in Education, Washington, D.C., March 1975.
12. Bale, R. E. *Comparison of pretest and cumulative results of an item bias study.* Paper presented at the meeting of the American Educational Research Association, Los Angeles, April 1981.
13. Rudner, L. M., Getson, P. R., & Knight, D. L. *The effect of various test and item properties on five approaches to biased item detection.* Paper presented at the meeting of the National Council on Measurement in Education, San Francisco, April 1979.
14. Burrill, L. E. Empirical comparison of item bias methods. In *Test item bias methodology: The state of the art.* The Johns Hopkins University National Symposium on Educational Research, Washington, D.C., November 1980.
15. Craig, R., & Ironson, G. *The validity and power of selected item bias techniques using an a priori classification of items.* Paper presented at the meeting of the American Educational Research Association, Los Angeles, April 1981.
16. Rudner, L. M., & Convey, J. J. *An evaluation of select approaches for biased item identification.* Paper presented at the meeting of the American Educational Research Association, Toronto, March 1978.
17. Shepard, L., Camilli, G., & Averill, M. *Comparison of six procedures for detecting test item bias using both internal and external ability criteria.* Paper presented at the meeting of the National Council on Measurement in Education, Boston, April 1980.
18. Scheuneman, J. A posteriori analyses of biased items. In *Test item bias methodology: The state of the art.* The Johns Hopkins University National Symposium on Educational Research, Washington, D.C., November 1980.
19. Tittle, C. K. Judgmental item bias methods. In *Test item bias methodology: The state of the art.* The Johns Hopkins University National Symposium on Educational Research, Washington, D.C., November 1980.

REFERENCES

American College Testing Program. *Assessing students on the way to college: Technical report for the ACT Assessment Program.* Iowa City, Ia.: Author, 1973.

Anastasi, A. *Psychological testing* (3rd ed.). New York: Macmillan, 1968.

Angoff, W. H., & Ford, S. F. Item-race interaction on a test of scholastic aptitude. *Journal of Educational Measurement,* 1973, *10,* 95–106.

Angoff, W. H., & Sharon, A. L. The evaluation of differences in test performance of two or more groups. *Educational and Psychological Measurement,* 1974, *34,* 807–816.

Bartlett, C. J., & O'Leary, B. S. A differential prediction model to moderate the effects of heterogeneous groups in personnel selection and classification. *Personnel Psychology,* 1969, *22,* 1–17.

Bersoff, D. N. Testing and the law. *American Psychologist,* 1981, *36,* 1047–1056.

Campbell, J. T., Crooks, L. A., Mahoney, M. H., & Rock, D. A. *An investigation of sources of bias in the prediction of job performance: A six-year study.* Princeton, N.J.: Educational Testing Service, 1973.

Cardall, C., & Coffman, W. E. *A method for comparing the performance of different groups on the items of a test* (Research Bulletin 64-61). Princeton, N.J.: Educational Testing Service, 1964.

Cleary, T. A. Test bias: Prediction of grades of Negro and white students in integrated colleges. *Journal of Educational Measurement,* 1968, *5,* 115–124.

Cleary, T. A., & Hilton, T. J. An investigation of item bias. *Educational and Psychological Measurement,* 1968, *28,* 61–75.

Cleary, T. A., Humphreys, L. G., Kendrick, S. A., & Wesman, A. Educational uses of tests with disadvantaged students. *American Psychologist,* 1975, *30,* 15–41.

Cole, N. S. Bias in selection. *Journal of Educational Measurement,* 1973, *10,* 237–255.

Cole, N. S., & Nitko, A. J. Measuring program effects. In R. A. Berk (Ed.), *Educational evaluation methodology: The state of the art.* Baltimore, Md.: Johns Hopkins University Press, 1981.

Cronbach, L. J. Test validation. In R. L. Thorndike (Ed.), *Educational measurement* (2nd ed.). Washington, D.C.: American Council on Education, 1971.

Cronbach, L. J. Equity in selection—Where psychometrics and political philosophy meet. *Journal of Educational Measurement,* 1976, *13,* 31–41.

Cronbach, L. J. Validity on parole: How can we go straight? In W. B. Schrader (Ed.), *New directions for testing and measurement: No. 5. Measuring achievement: Progress over a decade.* San Francisco: Jossey-Bass, 1980.

Darlington, R. B. Another look at "cultural fairness." *Journal of Educational Measurement,* 1971, *8,* 71–82.

Donlon, T. F., Hicks, M. M., & Wallmark, M. M. Sex differences in item responses on the Graduate Record Examination. *Applied Psychological Measurement,* 1980, *4,* 9–20.

Einhorn, H. J., & Bass, A. R. Methodological considerations relevant to discrimination in employment testing. *Psychological Bulletin,* 1971, *75,* 261–269.

Green, D. R. *Racial and ethnic bias in test construction.* Monterey: California Test Bureau, 1972.

Gross, A. L., & Su, W. Defining a "fair" or "unbiased" selection model: A question of utilities. *Journal of Applied Psychology,* 1975, *60,* 345–351.

Guion, R. M. Employment tests and discriminatory hiring. *Industrial Relations,* 1966, *5,* 20–37.

Humphreys, L. G., & Taber, T. Ability factors as a function of advantaged and disadvantaged groups. *Journal of Educational Measurement,* 1973, *10,* 107–115.

[illegible], P. A comparison of three approaches for determining item bias in cross-national testing (Doctoral dissertation, University of Pittsburgh, 1978). *Dissertation Abstracts International,* 1979, *40,* 2703A. (University Microfilms No. 79-24,700)

Ironson, G. H., & Subkoviak, M. J. A comparison of several methods of assessing item bias. *Journal of Educational Measurement,* 1979, *16,* 209–225.

Jensen, A. R. *Bias in mental testing.* New York: Free Press, 1980.

Kallingal, A. The prediction of grades for black and white students of Michigan State University. *Journal of Educational Measurement,* 1971, *8,* 263–265.

Kirkpatrick, J. J., Ewen, R. B., Barrett, R. S., & Katzell, R. A. *Testing and fair employment.* New York: New York University Press, 1968.

Lerner, B. The minimum competence testing movement: Social, scientific, and legal implications. *American Psychologist,* 1981, *36,* 1057–1066.

Linn, R. L. Fair test use in selection. *Review of Educational Research,* 1973, *43,* 139–161.

Linn, R. L. Test bias and the prediction of grades in law school. *Journal of Legal Education,* 1975, *27,* 293–323.

Linn, R. L., & Werts, C. E. Considerations for studies of test bias. *Journal of Educational Measurement,* 1971, *8,* 1–4.

Lord, F. M. A study of item bias using item characteristic curve [illegible] In Y. H. Poortinga (Ed.), *Basic problems in cross-cultural psychology.* Amsterdam: Swets & Zeitlinger, 1977.

Lord, F. M. *Applications of item response theory to practical testing problems.* Hillsdale, N.J.: Erlbaum, 1980.

[illegible] *American Psychologist,* [illegible].

Messick, S. The standard problem: Meaning and values in measurement and evaluation. *American Psychologist,* 1975, *30,* 955-966.

Novick, M. R. Federal guidelines and professional standards. *American Psychologist,* 1981, *36,* 1035-1046.

Petersen, N. S. An empirical investigation of three models of [illegible] (Doctoral dissertation, Florida State University, 1975). *Dissertation Abstracts International,* 1977, *37,* 5766A. (University Microfilms No. 77-[illegible])

Petersen, N. S., & Novick, M. R. An evaluation of some models for culture-fair selection. *Journal of Educational Measurement,* 1976, *13,* 3-29.

Reschly, D. J. Psychological testing in educational classification and placement. *American Psychologist,* 1981, *36,* 1094-1102.

Rudner, L. M., Getson, P. R., & Knight, D. L. Biased item detection techniques. *Journal of Educational Statistics,* 1980, *5,* 213-233.

Sawyer, R. L., Cole, N. S., & Cole, J. W. L. Utilities and the issue of fairness in a decision theoretic model for selection. *Journal of Educational Measurement,* 1976, *13,* 59-76.

Scheuneman, J. A method of assessing bias in test items. *Journal of Educational Measurement,* 1979, *16,* 143-152.

Schmidt, F. L., Berner, J. G., & Hunter, J. E. Racial differences in validity of employment tests: Reality or illusion? *Journal of Applied Psychology,* 1973, *58,* 5-9.

Schmidt, F. L., & Hunter, J. E. Racial and ethnic bias in psychological tests: Divergent implications of two definitions of test bias. *American Psychologist,* 1974, *29,* 1-8.

Stanley, J. C. Predicting college success of the educationally disadvantaged. *Science,* 1971, *171,* 640-647.

Stanley, J. C., & Porter, A. C. Correlation of Scholastic Aptitude Test scores with college grades for Negroes versus whites. *Journal of Educational Measurement,* 1967, *4,* 199-218.

Temp, G. Test bias: Validity of the SAT for blacks and whites in thirteen integrated institutions. *Journal of Educational Measurement,* 1971, *8,* 245-251.

Tenopyr, M. L. The realities of employment testing. *American Psychologist,* 1981, *36,* 1120-1127.

Thorndike, R. L. Concepts of culture-fairness. *Journal of Educational Measurement,* 1971, *8,* 63-70.

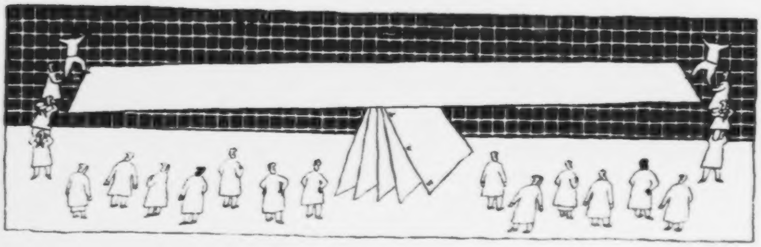
Veale, J. R., & Foreman, D. I. *Cultural validity of items and tests: A new approach.* Iowa City, Ia.: Westinghouse Learning Corporation/Measurement Research Center, SCORE Statistics Unit, 1975.

The SAT in a Diverse Society

Fairness and Sensitivity

The SAT undergoes meticulous checks to guard against ethnic or cultural bias.

by Thomas F. Donlon



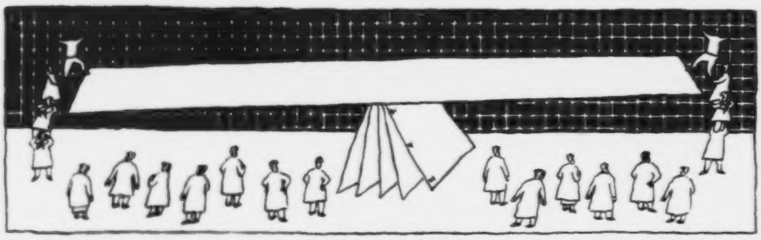
Reprinted from
THE COLLEGE BOARD
Review.
NO 122 WINTER 1981-82

by Thomas F. Donlon

The SAT in a Diverse Society

Fairness and Sensitivity

The SAT undergoes meticulous checks to guard against ethnic or cultural bias.



When the College Board introduced the Scholastic Aptitude Test (SAT) in 1926, one of the objectives was to provide colleges with assistance in coping with a growing diversity among their applicants. The Commission on Psychological Tests, a group of eminent psychologists, had been given responsibility for evaluating the suggestion that there be an SAT, and it began its report in 1926 by citing the changes in enrollment: "Statistics concerning higher education very plainly show the numerical increase of college population. . . . The natural consequence is that many institutions have sought to develop more adequate means for selecting from among the applicants those best fitted to profit by the opportunities offered." To provide "more adequate means," the committee recommended that there be an SAT. At this point, of course, the College Board had been in existence for over twenty-five years, and annually offered a number of achievement tests. But the Commission saw the limitations of the Achievement Tests, with their heavy dependence on curriculum, for these widening applications. "In some cases," they wrote, "limitations of educational opportunity would seem to be a factor in causing low marks in Board examinations. . . . This would be expected, since the Board examinations measure specific preparation. . . . a candidate whose educational opportunities have been limited has a much better chance . . . (on) a test which is not a measure of specific preparation . . ."

All in all, this fundamental premise for the SAT, that it offers ". . . a better chance" in the face of "limitations of educational opportunity," has been fulfilled. But as the world of college education widened, the SAT, because it was *not* a measure of specific preparation, because it reflected attainment in a very broad and general way, came to provide a meaningful common yardstick for facilitating the appraisals of an ever-expanding and increasingly diverse population.

But this general success in measuring varied groups did not blind the College Board to the possibility that there could be problems with the interpretation of SAT scores due to population diversity. Precautionary studies of the test performance of males and females, for example, were conducted from the very beginning, reflecting the strong concern on the part of the Board that there be no inappropriate differences in performance. In general, such studies demonstrated the appropriateness of the test for a wide variety of groups. The SAT, by virtue of the breadth of its coverage, and the careful editing of its content, is a balanced instrument with relevance for a variety of candidates. Although it is now only two hours long, it covers a range of topics and tasks that tends not to favor any one subgroup.

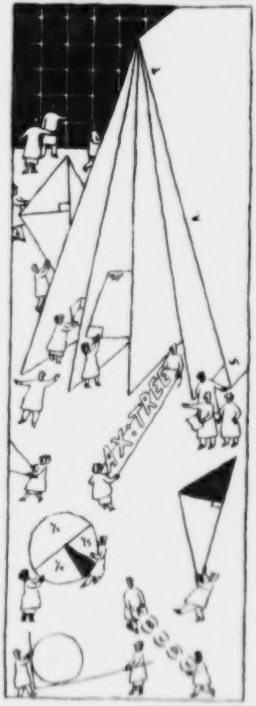
The questions of its fairness increased in frequency, however, with the emergence of a strong national concern for equity in access to higher education in the 1960s and 1970s. Is the SAT, in fact, unbiased? To what extent, for example, is it equally appropriate for men and for women? In general, the answers to such questions have been positive. As the college-going population has grown in numbers and broadened in variation, the SAT has continued to be appropriate for candidates with widely diverse backgrounds and curricular emphases. In recent years, the efforts to keep it that way have been intensified. However, while current evidence based on statistical comparisons of subgroups confirms that the test is basically fair, there is a need for a continuing program of review and analysis of the test from this standpoint. In the 1960s and 1970s a number of studies were undertaken, and since 1973 the types of questions used on the SAT have been periodically appraised for their appropriateness for different groups. This article provides a description of the procedures which are used in the test development process to ensure appropriateness, and the principal approaches which are used in the statistical analysis to evaluate test fairness.

Fairness in Content

There are several ways in which problems of fairness may arise regarding the test. Some of these are readily obvious, are reflected in clear imperfections in the content, and are apparent to a reader; others are hidden, detectable only in some characteristic of the scores. Thus, inappropriate content in a test may be outright offensive to certain groups, either through the portrayal of negative stereotypes, or through the diminishing of a group's importance through a failure to recognize it. Each of these flaws may create problems for test takers who notice the content defect and whose test performance is affected by it.

The more obvious problems of faulty content are readily avoided. Such avoidance requires a certain vigilance in editing, and a knowledge of subgroups and their reactions, but there is no special mystique to the process. Since its earliest days, the SAT has been carefully developed and reviewed so as to avoid material that may be offensive to anyone. The contemporary concern is merely an extension of this traditional process.

The effort to screen material is largely successful. Occasionally, a reading comprehension passage may generate concern on the part of someone who disagreed with it. Generally this has happened in the context of the so-called argumentative passage. Each form of the SAT from about 1950 on has had an argumentative passage, described in the specifications as "the representation of a definite bias on some subject," and often such passages present an impassioned argument for a fairly extreme position. Questions based on such passages are intended to test the candidate's ability to spot a specious argument and to deal with strongly opinionated material. In spite of disclaimers by the College Board or Educational Testing Service (ETS) of any approval of the opinions expressed in the test material, there may sometimes be a reaction from candidates, particularly from those who are opposed to the particular viewpoint expressed. Arguments against athletics, or democracy, or a graduated income tax do not upset the vast majority of the candidates, but the range of diversity among candidates is so great that some small number (believing strongly in athletics, or in democracy, or in a graduated income tax) may react with concern. The problem has not been unique to testing, of course; it pervades all of education in a society such as ours, in which a few people may feel very strongly about some things in a way that the vast majority does not. Granting the minority their rightful voice or influence is often a difficult matter. It is, however, an important problem and one which must be dealt with.



In general, such difficulties have been relatively minor. The test is conservatively edited, and it is not an instrument for social change. Throughout the years from 1926 to the late 1960s, consistent with the contemporary trends in educational text books and the media in general, the basic standard for appropriate test content was simply that it should reflect the mainstream of education and of life—the majority experience. While no overtly offensive or objectionable material was allowed to creep in, the content of the test, in sampling from mainstream prose, avoided direct reference to minorities or to minority-related problems.

Beginning in the 1960s, the prevailing treatment of minorities and women was widely challenged in society. The predominance of white male role models in the media and in the arts was viewed as overstated, and as inculcating expectations of sex and racial differences which worked to the detriment of women and of racial and ethnic minorities. Widespread changes began to appear in newspapers, in magazines, and in text books, as the language underwent a ripple of reform and as "Ms." began appearing in correspondence everywhere. Suddenly there was heightened awareness of the absence in the media of a balanced treatment of the sexes and racial and ethnic minorities. Reflecting these national patterns, the SAT began to change. The policy against overtly offensive content, content which could be upsetting to anyone, was, of course, retained. But a new, affirmative policy for the test emerged. Not only must negatives be avoided, in the sense of derogatory stereotypes, but there must be what one black educator called "respect signals," positive acknowledgement of the existence and accomplishments of various ethnic and racial groups and the diverse histories they reflect. A failure to deal openly with minorities would no longer do.

These forces were responded to, and the emerging patterns are reflected today in revised content specifications that require, for example, one minority-oriented passage in each form of the S.A.T., and an appropriate variety of references to women and minorities throughout the material. These changes appear most vividly in the reading comprehension passages and in the sentence completion items, which have more text than the analogy or antonym or other types of questions. The Test of Standard Written English also consists of material which can reflect cultural diversity, and it, too, is also carefully controlled in this manner.

These screenings of material from various viewpoints have come to be called "sensitivity" reviews. The new practices are, of course, not limited to College Board tests, but are applied to all tests that ETS prepares.

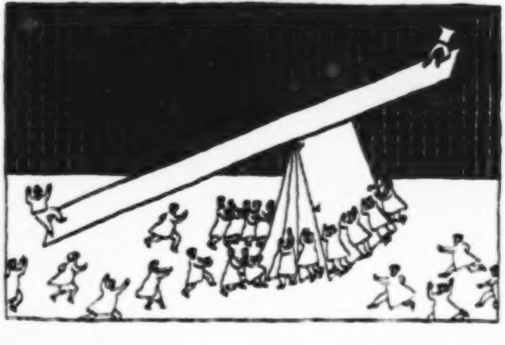
Sensitivity reviewers volunteer for their assignment. They are primarily test developers, since a knowledge of the subject matter areas covered in the tests is generally useful. Further, many reviewers are members of minority groups. Reviewing is not restricted to minorities, however. The guidelines clearly state that "it should be stressed that minority group membership is not a mandatory prerequisite to performing sensitivity reviews and serious consideration (is) given to *all* interested . . . staff who volunteer." What is stressed is not minority group membership *per se*, but the kind of sensitivity and awareness that can be developed through training, and that enables a reviewer to sense when material may be offensive. What's necessary is the ability to review tests from multiple perspectives, not simply from the viewpoint of a single subgroup or social/political perspective.

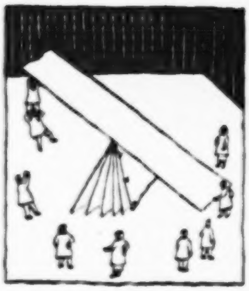
As this discussion suggests, the inclusion of certain material is as important, if not more so, than keeping other material out. For example, a question from the Test of Standard Written English might appear in either of two versions, as follows:

Version A. The newly enacted legislation
(A)
requires that all counties with
(B)
rural populations to provide
(C)
transportation to the polls and
absentee ballots. No error
(D) (E)

Version B. The newly enacted legislation
(A)
requires that all counties with
(B)
Spanish-speaking populations
to provide bilingual registra-
(C)
tion and election materials.
(D)
No error
(E)

It is obvious that the question still measures the same fundamental point about grammar, regardless of the content reference within which it is framed. The point is that specific content is often ancillary to some other purpose in designing a question, and that the modern goal of a test that is reflective of cultural diversity may often be achieved by adapting the material.





In some Achievement Tests, however, questions dealing with such matters as the migration of blacks from rural to urban communities, or with women's enrollment in courses in science, may be directly useful for measuring the outcome of instruction and study. Such minority-related questions will be included, but only after a careful review for appropriateness both from a cognitive dimension and an affective one. At the same time, they must be judged to meet the general standard that they are "both relevant and essential to effective measurement."

Direct questions of this type, calling for a knowledge of minority matters, are more often likely to occur in historical subjects, literature or literary subjects, legal studies, and psychological subjects. There can be no mechanical guidelines for determining the decision of what is to be included, and in what form. The policy guidelines for reviewers call for the exercise of "prudent judgment" and a consideration of material relative to "the context of the entire test." The policies are not policies of *exclusion*; they are policies that seek to minimize the potential for negative reaction to material and to require that all material be justified by some function that makes it necessary to use it in a test.

Statistical Checks

The new ways of reviewing test content are not based on the assumption that the score patterns will be different for any modified items. That is, changing a sentence completion question from a reference to Abraham Lincoln to one mentioning Susan B. Anthony or Martin Luther King does not usually alter the success rate on the question. The new policies are justified by *values*, rather than by statistics.

Statistics, however, are important, and the SAT is carefully studied from this viewpoint, also. The basic statistical facts that emerge from comparisons of scores are fairly well known within the educational community: among SAT candidates, males do substantially better than females on mathematical material, while white majority students of either sex do better on the SAT-verbal section than counterparts from such minority groups as Puerto Ricans, blacks, Mexican Americans, Oriental Americans, or Native Americans (Indians). On the SAT-mathematical section, Oriental Americans do best, followed by whites and other minorities. These patterns pose challenging problems to test sponsors who must show that they do not result from some flaw in the test material and that the test score differences reflect differences which will afford valid prediction.

The basic way to look at such score differences is to compare the predictive power of the SAT for the two sexes or for the several minorities. Using this approach, if the test is not biased, two candidates with the same score should perform about equally well in college, regardless of their subgroup membership. The number of studies of this type constitutes a fairly voluminous literature. In 1978 the College Board sponsored a summary by Breland* that considered the usefulness of SAT and of High School Record (HSR), among other measures, in a variety of studies of the college admissions process. In general, this survey supports the appropriateness of the SAT for many populations. It may sometimes be advisable to develop a special prediction equation for a given minority group, since SAT equations based primarily on white males may overpredict for blacks and underpredict for females. But the SAT offers predictive value for virtually every group it encounters.

Even though such studies of prediction tend to show no unfairness in total score on the SAT, the fact that there are subgroup differences in average score level cannot be ignored in a test that is widely used as one element in college admissions decisions. There must be an effort to *explain* this observed score difference, in order to promote test fairness. Accordingly, additional methods that consider the performance of groups on individual questions and clusters of questions, rather than on the test as a whole, are used. Are any questions inordinately hard or easy for certain groups? Are

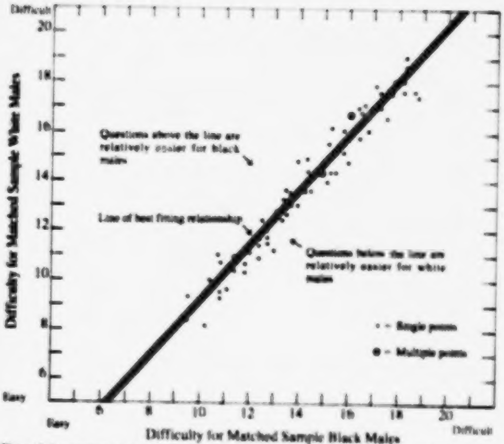


Figure 1. Comparative difficulty of SAT verbal questions (April 1975 form) for samples of white and black males approximately equal in verbal ability

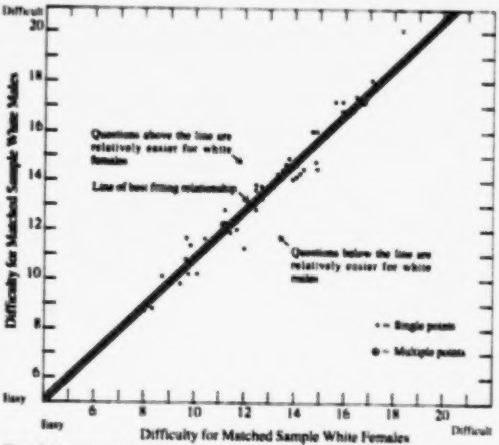


Figure 2. Comparative difficulty of SAT mathematical questions (April 1975 form) for samples of white males and females approximately equal in verbal ability

---

* Hunter M. Breland, *Population Validity and College Entrance Measures*. New York: College Entrance Examination Board, 1979.

The Science of Test Fairness: A Closer Look

One aspect of an analysis of differential difficulty is an examination of individual questions that seem to be farthest from the line of general tendency (or line of best fitting relationship in Figs. 1–2). The fact that they show *greater* than average distances tends to raise questions about their functioning. As mentioned in the text, such individual examinations have not generally been fruitful so far.

An example of a question which was relatively harder for blacks was.

RUNNER:MARATHON::

(A) envoy:embassy
(B) martyr:massacre
(C) oarsman:regatta*
(D) referee:tournament
(E) horse:stable

* Correct answer.

Although "Regattas" are less frequently associated with the minority experience, they are far from common for most of the majority, as well. The statistical analysis looks as follows:

Group	*Omit*	*A*	*B*	*C**	*D*	*E*
White sample	3(%)	7(%)	13(%)	53(%)	18(%)	6(%)
Black sample	8	7	12	22	31	21

* Correct answer.

When the percentage of candidates electing each of the wrong answers is compared, it appears on the surface that the question worked somewhat differently for the general groups because answer choices (D) and (E) together proved twice as attractive for the black sample as for the white. Is it possible that experiences with or the meanings of "referee:tournament," "horse:stable" are different for the two groups?

Before accepting such an hypothesis, however, it must be noted that the black sample and white sample in the previous comparison are not matched in ability. To a certain extent, then, the sets of percentages describe noncomparable groups. When the black general sample is contrasted with a more equivalent sample of whites, the following results appear:

Group	*Omit*	*A*	*B*	*C**	*D*	*E*
Matched white subsample	6(%)	7(%)	9(%)	26(%)	38(%)	16(%)
Black sample	8	7	12	22	31	21

* Correct answer.

It is clear that [illegible] matched whites perform very similarly to blacks. The clear implication of this is that it is not a black subculture which influences these patterns (for whites are not raised in that subculture), but some sort of general lack of knowledge or understanding of this item. Further confirmation of this is provided by comparing a subgroup of black candidates who are approximately equivalent in score level to the general white population.

Group	*Omit*	*A*	*B*	*C**	*D*	*E*
White sample	3(%)	7(%)	13(%)	53(%)	18(%)	6(%)
Matched black subsample	6	10	15	47	18	5

* Correct answer.

These blacks respond to the question more like whites of equal ability, in general, than like the black or white groups that are lower scoring. In short, there is no clearly apparent racial/ethnic content in this item, even though it appears on the surface to be relatively more difficult for blacks when random samples are used. When the responses are examined in depth, the patterns of success and error for racially defined groups of comparable ability are not different. What seems at first glance to be a possible case of cultural difference is seen upon closer inspection to be simply a result of differences associated with score levels. The question looks different to groups of different score levels, but it does so regardless of race.

These questions are typical. Another question on which minorities demonstrated relatively *favorable* performance was:

It is undeniable that Cervantes, though his work has ________ the age which it described, was a child of his time.

(A) enlarged
(B) nurtured
(C) delineated
(D) comprehended
(E) transcended*

* Correct answer.

Again, it is useful to adjust raw data by determining more comparable groups. The following table shows the performance of four groups: the general samples and two specifically selected subsamples, blacks matched to typical white candidates, and whites matched to typical black candidates.

Group	*Omit*	*A*	*B*	*C*	*D*	*E**
Matched black subsample	12(%)	8(%)	21(%)	9(%)	7(%)	43(%)
General white sample	8	11	19	9	12	42
General black sample	10	12	17	11	14	35
Matched white subsample	8	17	19	9	14	35

* Correct answer.

There is not a great deal of difference between the races on this question, even when the difficulty is not adjusted. When the adjustment is made, the similarity is strong. Distracter B is the most popular choice for all samples; the remaining choices are very evenly distributed. T.D. ●

there patterns of question content that might explain differences in difficulty for different subgroups? Total test score still figures in the analysis, but the methods presume that the test is, on the whole, unbiased.

The approach may be called "differential difficulty analysis," for it tests whether those questions that are difficult or easy for one group are the same ones which are difficult or easy for another. If they are not, if many questions shift position from hard to easy as they are administered to one group or the other, there is evidence that the test works differently for the two groups.

Probably the quickest way to describe the method is to resort to a diagram, as in Figure 1. The axis on the left side reflects the difficulty of the questions (items) for a white sample, whereas the line on the bottom reflects their difficulty for a black sample. The data are from an analysis of the form of the SAT-verbal section which was given in April 1975.

The numbers that describe the difficulty scale for the questions are called *deltas*, which can range from 5.0 to 21.0. They are based on the percentage of a group answering the question correctly. A delta of 13 is yielded when 50 percent of the group select the correct answer. Very difficult questions (10–20 percent pass) yield deltas of 18–21, very easy ones (80–90 percent pass) deltas of 5–8.

The samples of candidates used in this analysis were a random sample of black males and a sample of white males approximately matched on a verbal test. In general, results are clearer if the two groups under comparison are approximately equal in ability, as they are in this example. The points in Figure 1 indicate that questions vary in difficulty for the two groups in similar ways. Questions that are easy for one group tend to be easier for the other; questions that are harder for one group tend to be harder for the other. Most questions cluster closely about the line of best fitting relationship, which is not a statistical regression line but one that minimizes distances in both directions.

Another plot is shown in Figure 2. This shows the performance of white males and females on the 60 questions in the April 1975 form of the mathematical portion of the SAT. Differences in performance between the sexes on mathematical material have long been observed. Again, however, there is evidence of consistency of difficulty between the two groups.

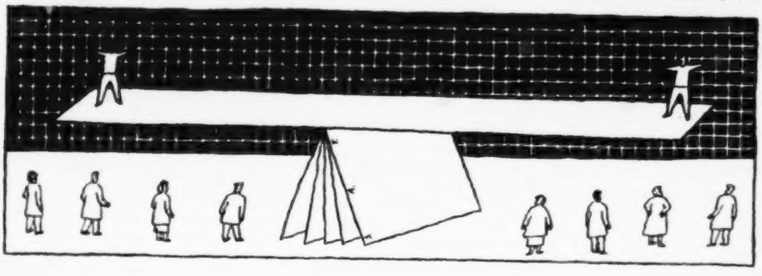
Figures 1 and 2 are fairly typical. In general, if a question is relatively harder for one group, it is relatively harder for the other. The items all cluster around the line of best fit, in a relatively narrow band.

The first step in any evaluation is to consider the basic overall information concerning the consistency of difficulty. How different is the white and black experience of the questions, as indicated by their relative success? Do the questions rank in the same order of difficulty for both groups? The typical answer to this question is a correlation coefficient, the statistical index that shows level of relationship from 0 (no relationship at all) to +1 (a consistent relationship). In educational and psychological literature, correlations of .95 to .99 are considered very high. But such correlations are, in fact, what the typical fairness analysis for the SAT shows. In six studies performed by ETS since 1973, the correlations of item difficulties between whites and blacks of both sexes were between .94 and .98.

It is important to note that what is being correlated in such studies are the two sets of question difficulties or deltas, one defined by white performance, one by black. The high correlations mean that the rank order of difficulty level for questions tends to be the same in the two groups: a question that is relatively harder for blacks is relatively harder for whites (relative, that is, to the other items). These very high values indicate that the test works in fundamentally consistent ways for both whites and blacks, and these numbers constitute the major finding of the research to date. They suggest that, whatever the factors are that affect score differences, they are not simply a case of poorly chosen question types or some source of inappropriate content. Questions are relatively hard or easy for the two groups in consistent ways.

But the questions do show some variation in their distance from the line of best fit. Some are virtually right on it, some above, some below. The second step in the differential difficulty method is to measure the distance between the point associated with a question and the line of best fit. Those questions which are farthest away are called "outliers," and they are selected for further study. They are the ones that show differences that would *not* be expected from the overall score differences for the groups. Thus, they are in a sense inconsistent with the total score data. The question naturally arises: As a group, do these questions have any characteristic in common that could explain why they are the farthest ones away?

Generally speaking, the answer to date, based on several studies, is "no." The position of a given question on such a chart is often due to sampling error, and the particular questions that emerge from a given study are often not the same ones that would emerge if the study was repeated with data from a different sample of respondents. An inspection of questions with large distance measures most often shows them to lack any rational characteristic which would explain their being "outliers." Some examples of



such questions accompany this discussion (see sidebar, page 5.) These are typical, and they are no more plausibly related in content to the stereotyped cultural differences than the rest of the questions in the test.

Not only may these distances be used to identify individual questions, they may also be averaged to compare the properties of different types of materials. In the SAT-verbal section, for example, the data in Table 1 emerge from analysis of the average distance of the four SAT-verbal question types. Table 1 summarizes the results of four SAT forms.

Table 1. Average Distance from Item to Line for Four SAT-verbal Question Types

Question type	*N*	*Average distance*	*Ranges*
Analogies	80	−.16	−1.29–+1.02
Sentence completion	60	−.12	−0.99–+0.80
Reading comprehension	100	+.08	−0.92–+1.17
Antonyms	100	+.11	−1.09–+0.94

A minus sign means that the items were, on the average, relatively harder for blacks, a plus sign that they were relatively easier for blacks. Because of the method, the values within a test will balance, so that if two item types show average differences in one direction, the other two will show differences in the other direction. The averages in Table 1 tend to indicate that analogies and sentence completions were on the average somewhat more difficult for blacks (compared to whites) than were antonyms or reading comprehension.

These results do *not* mean that analogies and sentence completion are "biased against minorities," or that antonyms and reading comprehension are "biased against whites." These are really very slight average differences, and no appreciable change in score patterns would emerge if the test were reconstituted entirely of antonyms or reading comprehension items. The results are possibly due to sampling differences, or to differences in the average difficulty level of the question types. In general, an item type cannot be considered "biased" on statistical grounds alone; there must be some knowledge of why the results are obtained.

The comparison of the item types in this manner is a demonstration of the general approach to the use of average distance measures. Using this approach, it is possible to study reading passages of different content, or mathematical questions which have diagrams associated with them, and so forth. The method enables the analyst to compare the average distance of a variety of interesting categories of questions.

One of the principal limits in an ordinary differential difficulty analysis, however, comes from the fact that many groups differ in total score level. This tends to make outliers out of questions that are more or less sensitive to differences in total score. One way around this difficulty is to divide the candidates into subgroups of approximately equal ability. In several studies, for example, the familiar College Board 200–800 score range has been broken up into 100 unit bands: 200–290, 300–390, 400–490, and so forth. The groups of candidates in each range are compared with respect to success on the questions. Using this approach for a question, it is possible to show an overall difference between the groups but to fail to show any significant difference at any of the smaller ranges considered. When this happens, it is evidence that the overall difference in performance on such a question is a result of differences in average score levels between the groups. About 50 percent of all questions showing overall group differences do, in fact, fail to show significance in such "range comparison" tests when they are subjected to them. The remaining questions tend to meet both criteria: their overall differences are sufficiently different to make them unusually distant, *and* there is evidence of statistically significant difference within at least one score range. For such questions the evidence indicates that some factor other than total score levels is influencing the result. Statistics, however, cannot tell us what that factor is. They simply provide a signal that the question must be carefully reviewed if it is to be used at all.

A useful adjunct to differential difficulty analysis, used only informally up to this point, is called "distractor" analysis. The gist of the approach is provided in the discussion of actual questions that have shown group differences (see sidebar). The wrong answers to multiple choice questions are called "distracters," and, as the discussion shows, ethnic, racial, or sex-defined groups of different composition can be meaningfully contrasted with respect to their patterns of response to these various alternatives. The method works best for samples of equal ability, because differences of ability can introduce "artificial" differences. (See the example of Runner:Marathon.) But a distracter analysis, if it shows similarity of patterns of responses, offers confirmation that the internal solution processes for the groups are reasonably equivalent. That is, if the patterns are similar, there is not one "minority" mental process and another "majority" process. Blacks who score well do so in ways similar to those of whites who score well; whites who do not score well do so in ways that are like those of blacks who do not score well. As the sample items in the sidebar show, it is possible for blacks and whites, when properly matched, to show very similar processes. To date, the use of these techniques for College Board tests has been limited to informal inquiries by test developers who seek to understand individual items. A more systematic use of the method, however, may emerge for the future.

The statistical analyses for test fairness also consider other aspects of the tests besides the comparison of the difficulty of questions. For example, differences in the characteristic work rate of various groups are often suggested as a source of score differences. In this view, minority students may run out of time to finish, leaving large numbers of questions unattempted. Accordingly, in doing an analysis, a careful check is made of the proportions of whites and blacks who complete the sections of the test. For a set of samples matched on ability, the average percentage of blacks completing an SAT-verbal section was about 9 percent less than the average percentage of whites. The average percentage of blacks completing an SAT-mathematical section was about 4 percent less than the average percentage of whites. These are not very large differences. Nor were there great differences in the number of blank questions at the end of a test. On the average, for SAT-verbal sections whites left 1.2 of 40–45 items blank; blacks 2.8 items. For SAT-mathematical, whites left 1.1 of 25–35 items blank, blacks 1.7. The differences between the racial groups, then, are not large by either yardstick, "percent completing the test" or "total items left blank." They suggest that while a somewhat greater number of blacks may move somewhat more slowly through the test, the differences are modest and not likely to produce sizable score effects.

The Continuing Effort

Both the review processes and the statistical approaches are continually under development. "Sensitivity" issues demand a current understanding of the viewpoints of the candidates and a constant effort to perceive the test from the perspective of individuals in the significant groups. On the statistical side, there is work underway on more powerful methods that apply mathematical models to the evaluation of differential item performance. These more powerful statistics are more useful than current methods when groups differ in ability, and it seems likely that these differences between groups are virtually unavoidable in day-to-day operations.

The continuing effort to produce a fair test has been rewarded so far with excellent reports from reviewers and analysts but the programs must continue to develop newer and better tools as they go along. Questions of fairness are a very important aspect of a test, and they must be asked repeatedly, for as the candidate group changes, the answers to questions of fairness may change. In the meantime, the evidence developed so far supports the conclusion that the SAT is not likely to be unfair to any group. ●

About the Author

Thomas F. Donlon is program research scientist at Educational Testing Service and works mainly in the area of College Board programs.

Mr. **Edwards**. Thank you very much, Miss Rigol.

We now will hear from Dr. Carol Dwyer, who is executive director for test development, at ETS in Princeton.

STATEMENT OF CAROL ANNE DWYER

Dr. **Dwyer**. Thank you, Mr. Chairman. Good morning.

I have worked at ETS for 15 years as a test developer. I am trained as a psychologist and consider myself a measurement specialist. I have had a lifelong interest in gender and achievement. I am very happy today to have this opportunity to talk to you about these issues.

I would like to start by saying some things about standardized tests. Historically, one of the main objectives of the development of standardized testing was to set comparable or standard tasks for everyone to demonstrate their knowledge or skills. Part of the aim of this was to give comparability. We have heard this morning already that a grade of "A" from one teacher doesn't necessarily have the same meaning as a grade of "A" from another.

But another part of this development was specifically to improve fairness. What standardized admissions tests replaced were criteria that are unacceptable by today's standards. For example, family connections, the possibility of large financial donations to an institution, one's religion, race or sex. Then, as now, I believe that the alternatives to standardized testing are very poor, indeed, and are more so for disadvantaged groups.

I would like to talk a little today also about women and their test scores, with a focus, as most of my predecessors have done, on college admission tests.

I think we would all agree that women and their roles in education are changing, and the tests, I believe, can give us some very important information about this. But what do tests tell us now about these changes? Dr. Rigol has already mentioned the changes in the group of people who choose to take the SAT over time. Women are now the majority of that group, approximately 40,000 more this year than men. But we also know that these women who choose to take the SAT are a less privileged group, academically and otherwise, than the men.

My written testimony has gone over some of the facts of women's scores relative to men, and women's scores now as opposed to in previous years. Dr. Rigol has also alluded to this. But the decline particularly in women's scores on the verbal test has been so much discussed recently that I would really like to say some things specifically about it.

I feel there are two principal reasons why we have a decline in women's verbal scores relative to men's. The first—and I believe it accounts for the large proportion of that change—is the population change that Dr. Rigol just mentioned. Women who before would not have aspired to college are now taking the SAT, and that's good news for those women and good news for women in general. But it is not good news in terms of the score average.

But, based on much broader evidence than the SAT, this score decline is just one part of a consistent trend in a much larger picture. A very important piece of work has just been completed by the psychologists and measurement experts, Dr. Janet Hyde and Dr. Marcia Linn of the University of Wisconsin and the University of California at Berkeley, respectively. They have completed a meta-analysis, which is a quantitative review of 165 studies of verbal ability and the patterns of sex differences within these studies. They reached the conclusion that there is no overall sex difference in verbal ability today, and they believe that that overall sex difference that previously existed disappeared around the year 1974. They can differentiate the studies before 1974 from those afterward. There was a small advantage in favor of women before 1974, which I think is mostly what sticks in people's minds, and today their conclusion is that this is now gone across all ages and across all types of tests.

Their finding, which I think is a very significant one in this field, is perfectly in accord with information that we have from admission testing programs, which are typically volunteer samples, such as the American College Testing Program and the ACT, but also information on the population in general that we have from very good sources such as the National Assessment of Educational Progress.

I should also say, almost parenthetically, that around this same period in the early seventies, in a number of achievement areas that we think of as being traditionally female, such as foreign language learning, women also lost their advantage to men. I think that is a significant piece of educational information. Frankly, we don't know why this occurred, or why the period around 1974 seems so significant. But it is clear that the answers to these questions have to go beyond the tests themselves because, for one thing, in many of these studies exactly the same, unchanged, tests were given before, during and after that magic period, so that it really does need to be something within the people taking the test or their education that has changed.

Mr. **Edwards**. What you're saying is it's a different population of women who started taking the test?

Dr. **Dwyer**. Yes, they are definitely a different population of women. But I also think that women changed their behavior. At the same time they stopped taking the less traditional stuff, they began taking more math, more science. They began taking more of the tests in these "nontraditional" areas. I think there is perhaps some shift in their attention, but I'm speculating here, on the basis of the data that has come to light recently.

But the question before us today, I think, is are these changes evidence of bias. In a word, I think the answer to that is no. In trying to convince you of this, I would like to talk about bias in tests as well as bias in test use—and others before me today have made that distinction. But I would really like to reinforce it because it is absolutely critical.

Differences in scores occur because people differ. The average weight of women and men, for example, as a group, differs in this country. That doesn't necessarily mean that the bathroom scales they stand on in the morning are biased—although some of us certainly wish that they were. But few would argue that we have complete educational equality today for women or for minorities.

Therefore, good tests should reveal these differences, where they exist.

It is very tempting to blame the tests for telling us things we don't want to hear. But yielding to that temptation is only going to lead us to gloss over real educational and social problems. Of course, tests themselves can be biased or unfair. But as test developers, we see sex bias as a very important component of our central concern, which is making tests valid. And by valid, all I mean is that a test is accomplishing its intended purpose.

Now, when we speak of bias in test use, the use of tests as distinct from the tests themselves can be biased, and I would like to give you just one example of this. If you were trying to select people for jobs assembling electronic components, and you used a spelling test to select people for these jobs, that would be a biased use of that test. On the other hand, the very same test showing the very same kind of group differences used to select secretaries might well be a very valid use of that same test.

These questions about how tests are used, as Dr. Cole stated earlier this morning, are not primarily technical or statistical questions. They are a matter of social values and logic and priorities. There are many ways to use tests to predict criteria such as college grades. But the consensus of measurement specialists is that a number of prediction systems may be technically sound in any given situation, but the right one to choose depends upon your goals and priorities, not upon the technicalities of that system itself.

Now, an organization like ETS cannot and should not be making these values decisions for institutions and other test users. But we do have a responsibility to make tests like the SAT as technically sound as possible, to provide technical assistance to the institutions and users of these tests, to make recommendations to them about how to use them, and to set standards for appropriate test use. We do these things.

Charges that the SAT cheats women by underpredicting their grades just demonstrate a fundamental misunderstanding of the role of the tests themselves in a prediction and selection system. The SAT, when properly used, is a valid predictor of college grades, for both men and women. And we must remember, above all, that the most difficult questions and decisions that have to be made about issues like college selection require values decisions that must be made, whether or not tests are used at all.

Now, the question here is what does ETS do about this when it makes tests. It's a very important question, and it is important in my everyday life as a test developer. Our whole worklife and the complicated system of producing a standardized test is aimed at improving validity and at eliminating bias. But there are particular aspects of this that I think are very directly related to the issue of bias that I would like to tell you about briefly today.

There is a process called a Test Sensitivity Review, wherein every question, before it ever appears before a student, is reviewed by specially trained reviewers, using documented criteria to eliminate offensiveness, inappropriate language, and stereotypes. For example, we would not have a reading test that portrayed women only in domestic roles when it mentioned them. I have brought with me today and would like to enter into the record a copy of the description and guidelines of ETS' sensitivity review process which is applied to every test question before it ever goes in front of a student.

We also approach this question statistically by analyzing results after the questions have been given to the students. We have a system that we are using that was recently developed for operational use, called differential item functioning, where we try to match individuals—say males and females, or Blacks and Whites—on their knowledge, and then examine each test question individually to see if, for these matched groups of people, the males and females, for example, they respond differently to the question in a way that is unfair—that is to say, an irrelevant difficulty that may be related to their race or sex. We feel that this combination of the statistical information, plus the judgmental information, is a very powerful guarantee, much more powerful than either alone, against bias.

Most of our tests are also developed in conjunction with committees, who are typically high school and college educators. Very careful attention is paid to the composition of these committees—by race, by sex, by geographical area, by type of institution, as well as by subject matter expertise.

I had intended to tell you that also the ETS staff who work on this, who are test developers, are two-thirds female and one-third male. But that point has already been made. However, I can't help saying that it is absolutely not true that the women do the grunge work and the men get to chose which questions go in. Believe me, I'm in a position to know that and everybody pitches in. So who chooses I think is proportionate to the representation of men and women in our small population.

Mr. EDWARDS. Salaries are appropriately equal, too?

Dr. DWYER. Yes, I believe they are. ETS administers its salaries in a very highly structured way. We have developed, with internal statistical expertise, a salary-equity model that is applied to every individual.

Mr. EDWARDS. Everybody knows what everybody gets?

Dr. DWYER. No, they don't. But personnel knows, and we do systematic analyses. When imbalances occur in the course of a year, salary adjustments are made. That is reviewed annually, I believe.

I would also just like to tell you that ETS not only participates very heavily in but has pioneered research and development in test bias. The contributions of our staff I believe are universally recognized in the field of measurement in this area.

The American Educational Research Association is meeting in Washington this week and there are literally dozens of sessions, research papers and open discussions, being held this week by ETS staff on matters directly relating to test equity. This is not at all unusual. I think you could probably go back 20 years and find the same kind of record of participation.

We are very public about the research we do and about sharing our data, and eager—I think some people would say over-eager—to tell the world about what it is we do.

I would like to finish by saying something about the alternatives to testing. We do sometimes hear alternatives to standardized test-

BEST COPY AVAILABLE

ing suggested, but more often this issue is ignored, as if our difficult decisions would vanish if tests ceased to exist. But I can't stress enough that it's important to remember that whether tests are used or not, would we still have these difficult decisions to make about people.

Testing takes place in a complex social setting and has recognized limitations, but I firmly believe that no better alternatives exist, including the option of not testing at all, which would allow race and sex bias to reenter the decision process, unexamined and unchecked.

The alternatives that are sometimes suggested, such as using grade point average alone, or letters of recommendation, or personal interviews and ratings, all have reliability problems, validity problems, and especially fairness problems that are worse than those of carefully developed and carefully used standardized tests. Just as important is that these alternatives do not lend themselves to public scrutiny in the way that tests do. The focus on personal qualities also has historically worked to the detriment of disadvantaged groups. Traditional out-group members, among which I personally would include women and racial and ethnic minorities, have benefited from situations that are highly structured. This is what I think of sometimes when I think about the salary-equity model at ETS, which I feel is a very equitable organization. When the criteria are clear, when people know what the rules of the game are, that is when the disadvantaged groups can get ahead.

In closing, I would repeat the advice of some of the other presenters today, that we try to focus on the causes of the difference we're seeing, rather than narrowly on the indicators of the differences, if we're really going to improve education and contributions to society of women and minorities.

Thank you very much.

[The statement of Carol Anne Dwyer, with attachments, follow:]

STATEMENT

OF

CAROL ANNE DWYER

Executive Director for Test Development
School and Higher Education Programs
Educational Testing Service
Princeton, New Jersey

before the

Subcommittee on Civil and Constitutional Rights
House Judiciary Committee
United States Congress

at a hearing on

Fairness in Standardized Tests

April 23, 1987

Good morning, Mr. Chairman and members of the subcommittee. My name is Carol Anne Dwyer. For the past fifteen years, I have worked at the Educational Testing Service as a developer of tests. ETS is a measurement and research organization headquartered in New Jersey. We are most widely known for our standardized admissions tests, including the Scholastic Aptitude Test (SAT), which we develop for the College Board, the Graduate Record Examination (GRE), the Graduate Management Admissions Test (GMAT) and the Test of English as a Foreign Language (TOEFL), which we also conduct for sponsoring boards. I am presently in charge of test development for elementary school, secondary school, and higher education testing programs.

I am a psychologist, a Fellow of the American Psychological Association, and a member of its Educational Psychology Division's Executive Board. I have also served on the Executive Council of the American Educational Research Association and have been Vice President for Measurement and Research Methodology with that organization.

My primary professional research interests, beginning with my doctoral dissertation at the University of California, Berkeley, have been the fairness of tests, the relationship between gender and achievement, and the interface of technology and social values. I have conducted training activities for AERA, APA, and other associations and institutions on bias in testing, and have chaired and served on numerous womens' committees for AERA and APA. I was one of the founders of AERA's Special Interest Group on Research on Women in Education.

Understanding bias, and knowing how to avoid it, is at the heart of what we do at ETS. Fairness is integral to the term "standardized." In every aspect of our work, from the development of questions, to the administration of tests, to the scoring of answer sheets, to the reporting of scores, and to the use of our tests in society, we are involved in the constant pursuit of equity. The contributions of ETS to the test bias literature over many decades show clearly that ETS is a leader in research and development in this field.

This morning, I would like to talk about four major issues concerning the fairness of tests. First, a word or two about why we have standardized tests; next, the question of "bias" on tests. Then I would like to share with you some of the recent trends in standardized test scores for females and minorities (which are often mistakenly assumed to be evidence of bias). Finally, I'll discuss admissions tests and what we do to ensure their fairness.

Why Standardized tests

Now, about standardized tests..... One of the primary purposes of developing standardized educational tests, which have a history in this country back to the past century, was to ensure the fair treatment of every test-taker. "Standardizing" means that each student is exposed to the same or equivalent tasks, administered under the same conditions, in the same amount of time, with scoring as objective as possible. These methods overcome problems that would otherwise exist in comparing students from different grades, schools, or areas. Without standardized tests, their performance could only be evaluated by different teachers using different methods, according to different criteria for success, and this would create questions about equivalency. For example, a "B" from one teacher in one classroom may indicate more knowledge than an "A" from a different teacher in a different classroom. Or the top class rank in one school may represent the same level of achievement as an average rank in another.

Standardization has been particularly helpful in the college admissions situation. Previous methods of selection were sometimes based upon such considerations as family ties to the college, the potential of large alumni gifts, and other criteria such as race, religion, and sex. Standardized tests became — and still are — a major vehicle for promoting equity in admissions and thus access to higher education for women and ethnic, racial, and religious minorities.

Standardized tests, along with high school grades, have proven useful to both students and colleges as an important element in affecting appropriate matches between them. Students benefit by their ability to select a college that will fit their academic preparation and expectations. Colleges make optimum use of their resources by admitting students whose test performance and high school record suggest that they are likely to be able to handle the work required and thus continue beyond the first year.

Decisions about selection and admission, however, are not the only reason for standardised tests. Uniform tests used by school systems or states provide helpful information that can lead to improved teaching and learning by pinpointing where deficiencies exist and where special efforts and funds should be targeted. Scores from repeated assessments of samples of a state's or the nation's students are also extremely valuable as indicators of educational trends. They provide some of the best and most useful information we have about what our students know and can do. Without these uniform tools, we would find it difficult to judge objectively whether boys and girls, or Blacks, Whites, and Hispanics across the nation, for example, perform the same or differently on important school tasks. We wouldn't know for sure what proportion of our young adults are literate, and we would have great difficulty determining whether our youth are prepared for the technological age and for the competitive world economy they face. Even if we guessed that our schools need reform — and were right — without standard measures, we would lack essential data for determining whether the reforms had worked.

Thus, there are extremely important reasons for having, and keeping, standardized tests in this country. The important issue that we are addressing today is the fairness of these tests. There is a great deal being said these days about bias in tests, and next I'd like to say a few words about that.

The Question of Bias

Some people think that a test is biased if different groups of people get different average scores. However, score differences in and of themselves do not mean that a test is biased; they may simply mean that the groups on average know different amounts about what is being tested. Measuring instruments that show differences are not necessarily biased. The average height of men, for example, is not the same as the average height of women, but this does not mean that yardsticks used to measure them are biased. Individual differences, whatever their source, are also recognised as inherent in the human condition. No two people are identical; no two groups are exactly alike.

In our educational system, individuals and groups differ in such respects as background, interests, quality of education they receive, types of courses taken, attitudes toward these subjects, kinds of non-school experiences, and school grades received. We expect these differences; we are enriched by the diversity that many of them bring to our culture. We are alerted by other differences to important problems to be solved. Tests are not intended to eliminate or disguise these differences; they are intended to identify them, if they exist, as accurately as possible, whether the results are judged to be positive or negative.

It is important to distinguish between test results that show differences, and the factors that cause the differences. Scales, for example, do not cause people to gain or lose weight. Tampering with the instruments to cover up differences is tempting, but dangerous and wrong. Tests are an easy target when they reveal unwanted or unexpected results, but they are the wrong target. Changing tests simply to hide differences in achievement could lead us to ignore real problems that should be addressed.

There are, of course, ways in which tests can be biased or unfair. Avoiding bias is central to a test-maker's main concern — that of developing a valid test. By "validity" I simply mean the extent to which the test accomplishes its intended purpose.

A test itself, for example, could conceivably contain questions that are unfair to a group of test-takers because of offensive language or inappropriate presentation of group members. It could also contain content that is not accurately representative of the ability being tested or questions that are poorly worded or unnecessarily confusing. It is extremely important that tests be free of such bias, and I will tell you later in my presentation what we at ETS do to ensure that our tests are fair in all respects.

It is also possible that a particular use of a test, rather than the test itself, may be biased. Use of a spelling test to select people for jobs that require no spelling — such as assembling electronic parts — is a biased use. That same test used to select secretaries may be perfectly appropriate — even if the average scores of the secretaries and the electronics assemblers are the same. Potential bias can also occur when test scores are used to predict performance on an inappropriate criterion measure (i.e., an outcome we would like to predict, such as class leadership or future income). This can occur if the criterion measure itself is invalid or biased for certain groups, for example. Tests can also simply be used for the wrong reason.

How tests are most equitably used in society is not primarily a technical or statistical question. Test makers have a responsibility to supply technical assistance, make recommendations, and set standards of good practice for the services they supply; but fair test use is a question of values that goes beyond the test itself and its makers. The purpose of testing and the best strategy for dealing with any group differences should be defined before any use is made of tests. If a stated policy goal is to increase the number of minority nurses on a hospital staff, for example, a racially balanced group of trainees might be selected from a pool of qualified applicants all of whom passed a nursing exam, rather than being selected simply in rank order of their test scores. Or if a college admissions staff's primary need is to predict as precisely as possible (without over- or under-prediction) the performance of a group of applicants' first year grades, they could use estimation procedures that will maximize that precision. Validity studies provide valuable information to help colleges in making decisions as to which technical procedures to use in their admissions practices to accomplish their goals. Charges that the SAT cheats women by under-predicting their performance are not supported by the facts. These charges are based on a misinterpretation of the role of the test in prediction and selection. The SAT, when used appropriately, is a valid predictor for both men and women.

Trends in Score Differences

Although differences in test scores of different groups of people do not in themselves mean a test is biased, it is nevertheless important to examine score differences carefully. They could point to an area of potential bias, warranting further investigation. They could also point to areas where curriculum or instructional change is needed. Let's take a look at some of these group differences.

Compelling evidence now exists of diminishing differences between men's and women's verbal test scores. This finding is based on results from a host of measures, and is not merely a function of performance on the SAT. A recently-completed, but not yet published meta-analysis by Janet Hyde and Marcia Linn of 165 studies (not including the SAT) reports that the long-observed tendency toward higher verbal performance for females (about .25 of a standard deviation) has nearly disappeared. The difference was evident prior to 1974, but from that time onward, no meaningful general sex differences in verbal performance have been shown to exist within any age group they studied.

It is extremely unlikely that this trend was a result of changes in tests, for a wide variety of tests show the same effect, and many of them had not been revised at all throughout the time period when scores changed. As many of us remember well, the early and mid-70's were a period of great social and educational change.

Since 1972, women's scores on the SAT verbal section have also declined in comparison to men's. In the years just prior to 1972, women scored between two and seven points (out of a total of 600 points) higher than men on the SAT verbal section. Now, however, women are scoring lower than men on that section by 11 points. A difference of about 50 points between men's and women's SAT math scores observed since the mid-70's still remains today. In a slightly older age group, the American women electing to take the Graduate Record Examination perform less well on average than men on its verbal section. However, we need to remember that students decide whether to take tests like the SAT, ACT and GRE. The nature of the group of people taking these tests has changed, as I will discuss later.

The best and most representative evidence of the true reading and writing achievement of all American men and women comes from the National Assessment of Educational Progress (NAEP). NAEP tells us that 17-year-old women continue to read and write better than men, although the margin of difference in reading achievement has become smaller over the years since 1975. Most of the decrease in the difference on the NAEP students' performance is accounted for by increases in men's scores, rather than a decline in the women's.

Studies now in progress show that one major cause of the decline in women's average scores on admission tests relative to men's are demographic changes in the self-selected group of people who take the tests. The most important of these is that many more women are now taking the SAT than ever before. Whereas women constituted only 44.5% of the test-takers in 1965, now at 52%, they have become the majority. This no doubt means that more women are aspiring to higher education. However, there is evidence that these women on the average are not as well prepared academically as the women who previously took the test. Therefore, their mean scores should not be expected to be as high as those of their predecessors. The net effect of this is that when the "new" group of women is included in the score average for all women, the average goes down. There has been no corresponding trend for young male high school graduates.

We are also investigating the possibility that changes in test content could have contributed to the decline in women's verbal scores. The amount of science reading in the SAT changed during the 1970's, for example; however our initial research does not indicate that the dates of these changes coincide with the dates of the observed score changes.

The ACT Assessment program is the other large college admission testing program that, like the SAT, tests over a million students each year. Users of the ACT and SAT tend to be clustered in certain regions, with those using the ACT concentrated primarily in the midwest and the southern region. The ACT Assessment tests college skills somewhat differently than the SAT, but the general trends in males' and females' scores are highly similar in both testing programs. ACT also has experienced a growth in the proportion of women taking the test and has also seen evidence that the women taking that test have had on the average fewer courses in math and science than the male ACT test-takers.

Many of the issues that I have discussed today with an emphasis on women are issues for racial and ethnic minority group members as well. We should also remember that these are not separate categories: very substantial numbers of test-takers are minority women.

Very often minority group members score lower on tests than the majority group. It is generally observed, for example, that Black test-takers, regardless of sex, score well below White test-takers on many educational tests. The magnitude of the difference between Black and White candidates' scores is larger than all but a very few gender differences. Hispanic test-takers as a group, tend to achieve scores somewhere between those of Blacks and Whites. Asian-American test-takers, as a group, excel in mathematics and science tests, but do less well than majority group members on verbal tests. Again, none of these differences in themselves indicates bias in the test, but may simply reveal continuing disparities in the education of minority students of all ages. For example, we know that Black and Hispanic students are less likely than White students to be enrolled in an academic program in high school.

These broad generalizations hold true on major admission tests such as the SAT and the ACT Assessment. However, there is some encouraging news. A number of statistics from admissions tests, large-scale longitudinal surveys, and the National Assessment of Educational Progress suggest that the gap between majority and minority group scores is narrowing, particularly in reading. Different tests show differences in the rate of this progress but the overall trend is clear.

Mathematics represents a special problem area for both women and Black test-takers as a group. Black students, like women, tend to take less coursework in mathematics than majority males and to be underrepresented in higher-level math courses. This is, not surprisingly, correlated with their mathematics test scores, and is an important area where further affirmative efforts to increase women's and minority group members' participation in mathematics and science activities are greatly needed in order to improve their academic and employment options.

Ensuring Fairness in Tests

As we have seen, differences in performance on standardized tests by different groups have long been observed and are closely monitored by educational researchers and testing companies. A necessary first step in investigating score differences is to examine the test itself for any possible bias. I want to take a little time now to talk about how we at ETS try to ensure that tests are fair.

Today we are focussing on standardized admissions tests. These tests are familiar to many of us because we or our children have taken them for entrance to college, graduate or professional school. These tests have been developed by specialized testing organizations which adhere to professional standards of quality and fairness. The most recent and comprehensive testing standards were jointly developed by the American Psychological Association, the American Educational Research Association, and the National Council on Measurement in Education. ETS is committed to continuing to meet these and all other applicable standards.

In addition, ETS, under the leadership of our president, Gregory Anrig, has attempted to go beyond the standards of the profession as a whole and has created its own standards for the quality and fairness of the tests we develop. These standards, which are set forth in this booklet, meet or exceed the general professional standards. Chairman Edwards, I request that a copy of these standards be inserted into the record of this hearing. In a further effort to address the dual goals of fairness and quality, ETS has established an accountability system of audits of all our testing programs. We have also invited numerous panels of distinguished educators and other specialists to critique our practices and to comment on them publicly.

We believe that our admissions tests are fair, as fair as anyone knows how to make them, and that they are fairer than alternatives such as interviews and letters of reference. Among the many steps taken to ensure the accuracy and quality of the tests we develop, two are especially important in ensuring racial and sex fairness: the "Sensitivity Review" and the "differential item functioning" process, which I would like to describe briefly.

First, every question in every test developed by ETS must undergo scrutiny by specially trained sensitivity reviewers who follow rigorous, documented criteria designed to identify questions that may be called biased because of inappropriate or offensive language or content. The reviewers also check to make sure that the test is appropriately balanced with respect to representation of people in different groups and in different roles. For example, we would consider it unacceptable to have a test of reading comprehension that showed women only in domestic roles. I would like to have a copy of an overview of our Sensitivity Review Guidelines included in the hearing record, Mr. Chairman.

Further, ETS has developed and is in the process of introducing operationally new statistical measures of potential bias, or "differential item functioning." The basic idea behind these statistics is that people who know approximately the same amount about the subject being tested by a question should have similar chances of answering it correctly, regardless of differences in their race, sex or ethnic background. The statistics therefore first match two groups of people in terms of their relevant knowledge and skill, then compare their performance on each test question. This gives us a measure of a test question's "differential difficulty." These statistics will thus help to identify differences in performance that may reflect potentially inappropriate characteristics of certain test questions. Such statistics will be used by all the major programs for which ETS develops tests. The combination of statistical analysis with thorough and detailed professional reviews of all questions provides a much stronger guarantee against potential bias than would either method used alone.

I should also mention that one of ETS's basic components in the test development process to ensure test validity is the use of committees of educators to plan and develop tests. These committees are composed of subject matter experts, usually teachers or university professors. The committees include women and men and minority and majority group members from all parts of the country, all types of educational institutions, and all specialities within their disciplines. They bring a broad perspective to the material included in our tests and help ensure its accuracy. These committees work with an ETS test development staff made up of 86 women and 46 men.

ETS has a long history of contributing to research on test fairness and making the data we collect available to other researchers. Three current studies, funded by the College Board, are particularly relevant to today's topic. The first is a complete content history of all the SAT tests administered from 1960 to 1987, telling us exactly what was tested on the SAT and when. We can then examine over the years whether content variations did or did not coincide with group score changes. [As mentioned earlier, the changes in test content in the 1970's do not appear to have coincided with the dates of observed score changes.] Another study will use the "differential item difficulty" technique that I just described to examine SAT verbal questions to see whether content factors (such as science contexts) are responsible for score differences for men and women who are otherwise comparable in their overall verbal reasoning skills. A third study will expand our knowledge of the demographic characteristics of the women and men who take the SAT and the relationship of these characteristics to their SAT scores.

Fairness is also important in how tests are used. It is the job of testing companies to produce the best tests possible from a technical point of view, and to provide interpretive material and sound technical assistance to their clients and users as they decide how to use test scores. Admissions test results, obviously, are intended to enhance the equity and efficiency of the college selection process. Decisions about the use of test scores by colleges do not occur in a value-free context and are not under the direct control of ETS or any other agency.

Institutional and societal priorities are brought to bear on statistical data. A better geographical mix of students, for example, may be desired in the new first-year class at a small college in a Great Plains state. A larger number of ethnic minority students might be sought by an institution in the Pacific northwest; or a large, predominantly female first-year class may be sought by a formerly all-male private college in New England which has recently decided to admit women. Each of these colleges will and should make its own value judgments, according to its own priorities, as to how to use test scores equitably in the admission process. This was the view taken by the National Academy of Sciences' Committee on Ability Testing in 1977, which put it better than I can:

> "Even recognizing the inherent difficulties, we believe that admissions officers have to exercise judgment, case by case, as, in fact, many now do. The goal should be to effect a delicate balance among the principles of selecting applicants who are likely to succeed in the program, of recognizing excellence and of increasing the presence of identifiable underrepresented subpopulations." (P. 196)

Mr. Chairman, in closing I would like to summarize the major points I have made today:

- carefully developed standardized tests are more fair than the available alternatives, which frequently rely on subjective personal judgments about groups and individuals;
- without tests we would lack basic information about how well educational programs are working — information that is essential if we are to focus our resources on educational improvements at the state and national level that will be most beneficial;
- score differences exist, but by themselves do not mean bias on tests; many factors contribute to such differences;
- ETS, a leader in research on testing and test bias, uses processes for developing standardized tests that are thorough, careful and designed to make our tests as fair as possible.

I thank you for the opportunity to speak to you today about an issue that is near and dear to my heart. I will be glad to answer any questions you and the committee may have.

CAROL ANNE DWYER

EXPERIENCE

1983-present	EDUCATIONAL TESTING SERVICE: Executive Director, School and Higher Education Programs (SHEP) Test Development, Administrative Head of combined Test Development area, including the following departments: Science, Languages, Legal Projects, Mathematics, Literature & Writing, Verbal, Reasoning & Measurement, Education, Social Studies, Developmental Mathematics and Reading.
1982-1983	EDUCATIONAL TESTING SERVICE: Deputy Director, School and Higher Education Programs (SHEP) Test Development; Director of Admissions Test Development. Deputy Administrative head of combined SHEP test development area, including Achievement & Certification Test Development and Admissions Test Development. Administrative head of Admissions Test Development, which includes the following groups: Mathematics, Literature & Writing, Verbal Aptitude, and Reasoning & Measurement. The Achievement & Certification Test Development area includes the following groups: Educational Processes & Developmental Skills, Science, and Languages.
1976-1982	EDUCATIONAL TESTING SERVICE: Test Development Group Head and Program Director, Elementary and Secondary School Programs Administrative head of test development unit of Elementary and Secondary School Programs Director of developmental and operational testing programs (Basic Skills Assessment Program, Bermuda Secondary School Certificate Programme, Delaware Assessment Program) Developer of subject matter examinations in psychology for the Graduate Record Examination Board (GRE) and the College Board Member of ETS advisory boards and committees for: Program research policies, candidate misconduct, program policies, norming studies, test analyses, women's affairs, statistical analyses of tests, item analysis procedures, use of cut-scores, personnel classification and compensation, regional office planning, prior review of research, controversial issues in testing

CAROL ANNE DWYER (continued)

1974-1979 Coadjutant Faculty: Child Psychology (undergraduate); Counseling and Testing (graduate) Trenton State College, Trenton, New Jersey

1972-1976 EDUCATIONAL TESTING SERVICE: Associate Examiner, Elementary and Secondary School Programs

Coordinator and primary developer for major testing programs (Secondary Schools Admissions Tests, Nova Scotia Educational Assessment, Oregon Statewide Assessment, Harrisburg (PA) Early Childhood Language Assessment project and workshops)

Instructor, Intensive Resident Courses (Programs of Continuing Education)

- Assessment Concerns in Early Education
- Assessment and Evaluation in Educational Planning
- Evaluation of Performance-Based Teacher Education
- Criterion-Referenced and Objectives-Referenced Measurement
- Reporting, Interpreting, and Using Test Results
- Criterion-referenced assessment of basic skills

Administrative director of Bermuda Secondary School Certificate Programme

Consulting:

- Cleveland Board of Education
- Madonna College
- National Institute of Education
- Maryland Department of Education (assessment for accountability)
- Delaware Department of Public Instruction
- Bermuda Ministry of Education (high school graduation requirements)
- Nemos Institute (needs assessment and Title IX)
- Wellesley Center for Research on Women
- National Catholic Educational Association (testing outcomes of religious education)
- Wisconsin Research and Development Center for Cognitive Learning--University of Wisconsin, Madison

1971-1972	School Psychologist	Murray School District, Dublin, CA
1970-1971	Clinical Psychology Practicum (intern)	McAuley Neuro-psychiatric Institute, St. Mary's Hospital, San Francisco, CA
1969-1971	Research Assistant	ETS Berkeley Regional Office, Berkeley, CA
1969-1970	Research Assistant	Dr. N. M. Lambert, University of California, Berkeley

CAROL ANNE DWYER (continued)

1969-1970	Psychological and educational testing	Head Start evaluation, Berkeley, Unified School District; Stanford University School Mathematics Study Group; Far West Laboratory for Educational Research & Development California School for the Deaf, Berkeley

EDUCATION

University of Chicago	1976		Industrial Relations Center, Summer Management Development Seminar
University of California, Berkeley	1972	Ph.D.	Educational Psychology
University of California, Berkeley	1970	M.A.	Educational Psychology
Barnard College, Columbia University	1968	A.B.	Psychology

SELECTED PUBLICATIONS

Sex Equity from Early Education through Postsecondary. In *Achieving Sex Equity through Education*, S. Klein (Ed.). Baltimore: John Hopkins University Press, 1985.

AERA Guidelines for Eliminating Race and Sex Bias in Educational Research and Evaluation. *Educational Researcher*, 1985, *14*, 16-17.

Technology and Testing: Implications for Validity in the Computer Era. *Educational Measurement: Issues and Practices* (in press). Original version presented at Fifth International Symposium on Educational Testing, University of Stirling.

Equating the Standards of Educational Examinations in Two Countries. *British Journal of Educational Psychology* (with R.J.L. Murphy, in press).

Sex Bias and Reading Tests. Paper presented to the International Reading Association Annual Meeting, May, 1984. (with E. Gerritz)

Chair and presenter, APA Division 15 Recent Scholars Awards. American Psycholgoical Association Annual Meeting, Anaheim (CA), 1983.

Equity in a Cold Climate. *Educational Researcher*, 1983, *12*, 14-17. With S.K. Biklen, L.S. Koester, D. Pollard, J.D. Scheuneman, C. Shakeshaft.

CAROL ANNE DWYER (continued)

Encouraging Girls and Women in Mathematics (book review) *The Psychology of Women Quarterly*, 1983, *7*, 385-387.

Achievement Testing. Invited review of achievement testing for *Encyclopedia of Educational Research (Fifth edition)*. New York: Macmillan and Free Press 1982.

Review of J. Stockard, P. A. Schmuck, K. Kempner, P. Williams, S. K. Edson, and M. A. Smith *Sex Equity in Education Contemporary Education*, 1982, *1*.

AERA Invited Training Session: Bias and testing. AERA 1982 annual meeting (with J. Scheuneman).

Organizer and Chair, invited State of the Art series, American Psychological Association annual meeting August, 1982 (with Anne Anastasi, Robert Ebel, and Samuel Messick).

Assessment of Young Children. Invited workshop, International Council of Psychologists, University of Southampton (England), 1982. With W. M. McPeek.

The Role of Schools in Developing Sex Roles Attitudes Chapter 12, in J. Downing, et al (Eds.) *Sex Role Attitudes and Cultural Change*. Dordecht, Holland: D. Reidel, 1981.

Test development for adaptive testing. *Proceedings of the 23th Annual Conference of the Military Testing Association*, 1981, Vol. II, 1301-1312.

Training and Employment Experiences of Educational Psychologists. Paper presented to Northeast Educational Research Association, October 1980 (with Janice Scheuneman).

Equating the Standards of Educational Examinations in Two Countries. Paper presented to the Fourth International Symposium on Educational Testing, Antwerp, Belgium, June 1980 (with R. J. L. Murphy).

Criterion-Referenced Testng. Invited workshop, International Council of Psychologists, University of Bergen (Norway), 1980. With W. M. McPeek.

Validation of Performance Standards. Paper presented to the Fourth International Symposium on Educational Testing, Antwerp, Belgium, June 1980 (with C.L. Wild).

Sex bias in selection. In L. J. Th. van de Kamp, W. F. Langerak, and D. N. M. de Gruijter (Eds.), *Psychometrics for Educational Debates*. Chichester, England: John Wiley & Sons, Ltd, 1980 (With C. L. Wild)

The Role of tests and their content in producing apparent sex-related differences. In A. C. Petersen and M. A. Wittig (Eds.) *The Development of Sex-Related Differences in Cognitive Functioning*. New York: Academic Press, 1979.

CAROL ANNE DWYER (continued)

The role of schools in developing sex-role attitudes. Paper presented to the World Congress on Mental Health, Salzburg (Austria), July 1979.

Minimum Competency Testing: Problems and Solutions for the Eighties. Symposium on Minimum Competency Testing, Temple University: Philadelphia, October 1979.

Setting defensible performance standards (workshop). Phi Delta Kappa leadership Conference: Cleveland, Ohio, March 1979.

Minimal Competency Testing and Measurement Technology. *Interchange*, 1978, *3*, whole issue.

A cross-national survey of cultural expectations and sex role standards in reading. *Journal of Research in Reading*, 1979, *2*, 8-23. (With J. Downing)

A debate on the proposition: adequate measurement technology exists to implement fair, equitable, and useful minimum competency testing programs. In Center for Applied Performance Testing, *Proceedings of the National Conferences on Minimum Competency Testing*. Portland, OR: CAPT, 1978.

Sex bias in selection procedures and selection instruments. Paper presented at the Third International Symposium on Educational Testing, Leyden (Netherlands), July 1977. (With C. L. Wild)

Test content in mathematics and science: The consideration of sex. Paper presented at the American Educational Research Association, April 1976.

Test content and sex differnces in reading. *The Reading Teacher*, 1976, *29*, 753-757.

Test Content and the determination of sex differences. Paper presented at American Educational Research Association, Washington, DC., April 1975.

Sex differences in reading: A symposium. (Ed.) Washington, D.C., National Foundation for the Improvement of Education, 1975.

Comparative aspects of sex differences in reading. In D. Moyle (Ed.). *Reading: What of the future*? London: Ward Lock, 1975.

Comparative aspects of sex differences in reading. Paper presented at United Kingdom Reading Association. Ormskirk, England, July-August 1974.

The influence of children's sex-role standards on reading and arithmetic achievement. *Journal of Educational Psychology*, 1974, *66*, 811-816.

Sex differences in reading: An evaluation and a critique of current theories. *Review of Educational Research*, 1973, *43*, 455-467.

CAROL ANNE DWYER (continued)

EDITORIAL CONSULTANT

American Educational Research Journal
Educational Researcher
Encyclopedia of Educational Research (fifth edition)
Journal of Educational Psychology
Mental Measurements Yearbook
Quarterly Review of Development
Review of Educational Research
Educational Psychologist
Journal of Reading Behavior
Journal of Research in Mathematical Education

PROFESSIONAL ASSOCIATIONS

- American Educational Research Association
  - Member-at-large of the Executive Council, 1982-1985
  - Vice President of AERA for Division D: Measurement and Research Methodology, 1978-80
  - Committee on Long-Range Planning 1984-85
  - Chair, Standing Committee on the Status of Women, 1980-1982
  - Chair, Committee on Research Guidelines, 1980-1985
  - Reviewer Divisions B, C, D and H programs
  - Judge, Divisions D and H research awards
  - SIG Research on Women in Education--Program Chair 1974-75, Assistant Chair 1975-76, Chair 1976-77
  - Consulting Editor, *Encyclopedia of Educational Research*, fifth edition
  - Women Educators, Research Award Competition Judge (1980-1981, 1981-1982)

- American Psychological Association
  - Division 15 Continuing Education Committee, 1982-1984; Chair, 1984-1986.
  - Division 35, Program Committee, 1982
  - Division 15 Program Committee, 1981-1982, 1982-1983
  - Division 15 Nominating Committee (chair) 1981
  - Division 15 Committee on Women & Minorities in Educational Psychology 1979-1981
  - Research Guidelines Committee 1975-1976

- International Reading Association (National Committee Member 1975-1977, Sexism and Reading)

- National Council on Measurement in Education
  - Program reviewer
- International Association for Applied Psychology
- International Council of Psychologists

Honors

Fellow of the American Psychological Association

ETS SENSITIVITY REVIEW PROCESS

An Overview

Educational Testing Service • Princeton, New Jersey

Acknowledgment

The original procedures for the ETS sensitivity review process were developed by Ronald V. Hunter and Carole D. Slaughter.

Substantial contributions to the process have been made by other writers of earlier documents dealing with the issue of sensitivity. Many of these pioneering efforts, such as the *ETS Guidelines for Testing Minorities* and the *ETS Guidelines for Sex Fairness in Tests and Testing Programs*, provided much of the creative thought and detail contained within this document.

Finally, many ETS staff members have taken the time to review drafts of this document. In so doing they have provided a wealth of helpful suggestions and productive insights on this complex issue.

Copyright © 1987 by Educational Testing Service.
All rights reserved.

Educational Testing Service, *ETS*, and are registered trademarks of Educational Testing Service.

Educational Testing Service is an equal opportunity/affirmative action employer.

Table of Contents

Introduction 4

Background 4

Factors Guiding the Sensitivity Review Process 5
- Cultural Diversity 5
- Diversity of Background Among Test Takers 5
- Force of Language 5
- Changing Roles 6

The Sensitivity Review Process 6
- Reviewers 6
- Test Sensitivity Review Procedures 6
  - (1) Preliminary review 6
  - (2) Final review 6
  - (3) Arbitration 7
- Sensitivity Review Procedures for Other Publications 7
- Review Criteria 7
  - (1) Stereotyping 7
  - (2) Examinee perspective 7
  - (3) Underlying assumptions 7
  - (4) Controversial material 7
  - (5) Contextual considerations 8
    - Historical domain 8
    - Literary domain 8
    - Legal domain 8
    - Health domain 8
  - (6) Elitism, Ethnocentricity, and Related Problems 8

Additional Information 8

THE ETS SENSITIVITY REVIEW PROCESS:

An Overview

Introduction

Educational Testing Service is committed to the development of tests and other publications that reflect a thoughtful and humanistic consideration of all people and that acknowledge the multicultural nature of our society. In the 1970s, ETS broadened the review of all tests to ensure that: 1) they contained questions recognizing the varied contributions that minority members have made to our society and 2) there was no inappropriate or offensive material in the tests. In 1980, the corporation, building on the review procedures, formally adopted the ETS Test Sensitivity Review Process. In 1986, this process was extended to all publications, including audiovisual materials and art work. The purpose of the process is to ensure that the guidelines, found in the ETS Standards for Quality and Fairness, are met.

One such test development guideline instructs test developers to prepare for each test, with appropriate advice and review, specifications that cover several critical areas, including requirements for material reflecting the cultural background and contributions of major population subgroups.

Another test development guideline requires the review of individual items, the test as a whole, and descriptive materials to assure, among other things, that language, symbols, words, phrases, and content that are generally regarded as sexist, racist, or otherwise potentially offensive, inappropriate, or negative toward major subgroups are eliminated.

Finally, an accountability guideline demands the review of publications and other materials to eliminate language or material generally regarded as sexist, racist, or otherwise offensive or inappropriate.

Although a substantial portion of the process consists of general criteria that can be applied to any population group, experience has shown that a particularly vigilant effort must be made to evaluate our publications from the perspectives of the following groups: Asian/Pacific Island Americans, Black Americans, Hispanic Americans, individuals with disabilities, Native Americans, and women. The process, therefore, specifically addresses areas of special concern to these population groups.

Background

Sensitivity review, required by Educational Testing Service for all its tests and publications, attempts to eliminate offensiveness from all ETS materials. Such offensiveness could obstruct the intent of a publication -- whether a general publication or a test. In the area of test development, for example, the impetus to avoid offensive material comes from a desire to ensure that each test is indeed asking all test takers to perform the same task under the same conditions, insofar as it is possible to do so.

The importance attached to sensitivity review does not imply a measurable relationship between material considered offensive by some test takers and the scores of test takers. However, material that candidates consider offensive may produce negative feelings that may affect their attitudes toward tests, and hence, their test scores. Recognizing both the negative feelings that a test taker may have when dealing with test material and the possible effect that offensive test material may have on the test taker's performance, ETS has instituted a sensitivity review process for tests and other publications.

The sensitivity review guidelines specify six groups that are to be given special consideration in sensitivity review: Asian/Pacific Island Americans, Black Americans, Hispanic Americans, individuals with disabilities, Native Americans/American Indians, and women. The guidelines, however, are general; they can be, and are, extended to cover materials that are potentially offensive to the elderly and to members of other groups, including men, not specifically mentioned in the guidelines.

The sensitivity review promotes a general awareness of and a response to:

- the cultural diversity of the United States;
- the contributions of the various ethnic and minority groups and women to the history and culture of the United States as well as the achievements of individuals within these groups;
- the diversity of background, cultural tradition, and viewpoints to be found in the test-taking population;
- the force of language in setting or changing attitudes toward various groups and toward women; and
- changing roles and attitudes in United States society.

Factors Guiding the Sensitivity Review Process

Cultural Diversity

Since the 1960s, the United States has become much more aware of the diversity of its population. Both the civil rights and feminist movements have helped increase the visibility of women and people from minority groups. Further, this representation has moved away from stereotypes and has emphasized the occupational diversity and cultural contributions made by all groups.

Consistent with these advances in society as a whole, the ETS sensitivity review guidelines specify that all ETS publications must include material that reflects the diversity of the test-taking population. By underscoring the contributions of all groups to United States history and culture and by highlighting the individual achievements of women and minority groups in fields such as science, literature, and business, ETS tests and publications attempt to maintain a balance that acknowledges the cultural diversity of the test-taking populations. The sensitivity review process requires the demonstration of such a balance.

Diversity of Background Among Test Takers

Because test takers are different, a question may carry an emotional charge for one candidate or group of candidates that it does not carry for others. For example, a reading passage on sex differences in intellectual ability, a question on the problems of living in a ghetto, or data concerning the presence of certain diseases in a given population may very well be upsetting to some test takers. The sensitivity review helps to ensure that material dealing with disabilities, gender, or ethnicity is developed with care. Further, test takers may go away from a standardized test not knowing that they have given an incorrect answer or that they have misread a passage; therefore, offensive statements included as choices for the answer to a question may well reinforce the very stereotypes or bias that the rest of the test avoids. Such choices must be avoided wherever possible.

Force of Language

With changing attitudes toward various groups within the United States have come changes in the words we use. *Negro*, for example, is no longer generally acceptable as a racial group description; *Black American* is now the preferred term. At one time, people with disabilities were universally referred to as "handicapped." The term used most frequently now is *disabled*. A term such as "settlers and their wives" is no longer used because it places women in a category apart from settlers, who are generally considered male in this construction, and because it downgrades women's contributions to settlement. Similarly, the so-called "generic he," though at one time considered the correct pronoun to use when referring to both sexes, is now seen as excluding women. These and other words and descriptions that exclude groups or perpetuate stereotypes are avoided in ETS tests and publications.

Changing Roles

Significant social changes have taken place in the United States in recent years: Family patterns have changed; women have entered the paid labor force in greater numbers and in positions they have not typically held; members of minority groups are making important contributions to fields from which they were largely excluded just a short time ago. ETS tests and publications reflect such changes, indicating to test takers that ETS is aware of social change and of the opportunities open to all test takers. In ETS materials, therefore, job titles that seem to restrict occupations (*firemen, businessmen, stuntmen*) are not used. Further, women and members of minority groups are portrayed as active participants in society and appear in a balanced variety of roles. Where a question in a mathematics test might once have mentioned Mary Smith's calculations for roasting a turkey, a similar question today might mention her calculations for establishing missile trajectories.

The Sensitivity Review Process

Reviewers

Reviews of ETS publications are conducted by ETS professional staff members who are trained in sensitivity issues at two-day workshops and periodic one-day refresher courses. While there are a number of reviewers who are women and/or members of minority groups, membership in such groups is not a prerequisite; and any professional interested in the process and showing concern for equity may be trained to administer it.

Test Sensitivity Review Procedures

The test sensitivity review process has three components: an optional preliminary review (required by some testing programs), a mandatory final review, and an arbitration process.

(1) Preliminary review

Any staff member who is assembling a test may request a preliminary review to screen questions and answers, reading passages, and other materials for sensitivity-related issues. The reviewer's recommendations are not binding at this stage; however, a preliminary review is an excellent means of identifying potential problems early in the test development process, when modifications can be made more easily.

(2) Final review

The mandatory final review takes place after the test has been assembled and during the regular editorial process. This review must be conducted, even if the test received a preliminary review.

The sensitivity reviewer, who is always someone other than the person who is responsible for the test (the test assembler), notifies the test assembler in writing of any sensitivity-related issues the test has raised. The test assembler must then address in writing all concerns of the sensitivity reviewer. In the vast majority of cases, the test assembler and the reviewer are able to resolve the issues satisfactorily. When the two cannot resolve issues raised by the reviewer, a sensitivity review coordinator meets with them to ensure that they clearly understand each other's position. If the reviewer and assembler still cannot reconcile their differences, they and the coordinator meet with a test development director, and the four of them discuss the problem question or passage. Most issues are resolved at this point. In a few cases, the material in question must go to arbitration.

(3) Arbitration

Arbitration is performed by a panel of three staff members who are outside the test development areas and who are not involved with the test in which the disputed question or passage appears.

After examining the disputed material, the panel must reach consensus as to whether or not the material conforms to ETS sensitivity review guidelines and procedures. The decision of the arbitration panel is binding.

Sensitivity Review Procedures for Other Publications

Sensitivity reviews of ETS publications other than tests are performed by the editors of those publications unless the editor is also the author, in which case another editor performs the sensitivity review. Editors, like test reviewers, are trained in the sensitivity process.

As a rule, editors undertake sensitivity reviews when the manuscript has reached final draft stage, before it is put into production. However, editors are encouraged to review copy informally as early in the editorial process as possible. If a manuscript that has already received a sensitivity review is changed, the sensitivity review editor must review the additions for conformity to the ETS sensitivity guidelines. Editors are also responsible for reviewing audiovisual publications and artwork proposed for inclusion in publications, using the same procedures described above. ETS-developed software is also reviewed for sensitivity.

Editorial staff bring sensitivity issues to the attention of the project director. The editor then works with the project director to eliminate questionable or inappropriate material from the publication.

A project director who chooses not to change a manuscript must reply in writing to the editor's query. In case of further disagreement, the dispute is resolved with the same arbitration process as that used for test material.

Review Criteria

The sensitivity review training sessions teach reviewers to evaluate material in light of specific criteria:

(1) Stereotyping

All ETS publications are reviewed to ensure that their language and illustrations reflect a fair and unbiased attitude toward all people and are free of material that reinforces stereotypes. For example, women should not be portrayed only cooking, maintaining a home, or taking care of children. Sensitivity reviewers are trained to identify stereotypes specific to each of the targeted groups and are given a list of "caution words and phrases." Some of these are unacceptable, e.g., "redmen" when referring to Native Americans. Most caution words and phrases (e.g., *underprivileged*) signal that a sensitive issue is being addressed.

(2) Examinee perspective

Test sensitivity reviewers have a particular concern that does not apply often to reviewers of other kinds of publications. They must evaluate all questions from the perspective of test takers, who do not necessarily know the correct answers. If an examinee must know the correct answer in order to prevent a question from reinforcing negative attitudes or stereotypes, the question may be in violation of the guidelines. For example, a wrong answer to a question about Hispanic culture should not reinforce—for those who mistakenly think the answer is right—the stereotype of the "lazy" Hispanic who always puts off work until "mañana."

(3) Underlying assumptions

While stereotypes are often blatant, underlying assumptions can be extremely subtle. Underlying assumptions may lead one to mistake aspects of Western culture for universal norms or to misunderstand a particular group. For instance, a publication that refers to an "afflicted" person "suffering from" cerebral palsy reflects the writer's underlying assumptions about what it is like to have this physical condition.

(4) Controversial material

Highly controversial material, such as legalized abortion, is to be included in tests only when it is relevant to what is being tested. For example, a test for doctors or nurses may have to contain questions on abortion, but a test of reading ability should not include a reading passage on this controversial subject.

(5) Contextual considerations

Sometimes the use of potentially sensitive material is unavoidable. There are four main areas in which this may occur:

- *Historical domain*: In order to measure an individual's knowledge of history, it may sometimes be necessary to quote from material written during a period when social values differed markedly from today's. For example, an older passage describing members of the Black community may use the term "colored." While it is desirable to avoid such material when possible, the material must be judged in the overall context in which it appears.
- *Literary domain*: Material that is designed to measure an individual's knowledge of literature or quotes from works of literature often contains similar problems. For example, a passage may use the so-called "generic he" in referring to men and women. Again, such material must be evaluated in light of the overall purpose of the test.
- *Legal domain*: Material drawn from legal sources may sometimes deal with sensitive issues. For example, a law test question on the detention of citizens may refer to the incarceration of Japanese Americans during World War II.
- *Health domain*: Certain examinations in the health profession require knowledge that may be considered sensitive in other contexts. For example, it may be necessary to test nursing candidates' knowledge of Tay-Sachs disease in Jewish families.

Inclusion of potentially sensitive material depends on the content of the entire test or publication. Given an appropriate context, use of certain material may be justifiable.

(6) Elitism, Ethnocentricity, and Related Problems

To eliminate concepts, words, phrases, or examples that may upset or otherwise disadvantage a test taker, ETS makes every effort not to include expressions that might be more familiar to members of a particular social class or ethnic group than the general population, such as "soul food" and "trust fund," unless the terms are defined or knowledge of them is relevant to the purpose of the test. Words and sentence constructions that could have different meanings for different ethnic or geographic groups are avoided. Care is also taken to assess the appropriateness of dialect, slang, and non-English words and phrases, such as "bairn," "stickball," and "maven," which tend to be more familiar to certain ethnic, geographic, or other subgroups of English speakers.

Additional Information

The above is an overview of the sensitivity review process. If you have comments, questions, or desire more information about the process, please write to the Office of Quality Assurance, 09-D, Educational Testing Service, Princeton, NJ 08541-0001.

ETS STANDARDS FOR QUALITY AND FAIRNESS

Adopted by the Board of Trustees

Educational Testing Service • Princeton, New Jersey

PREFACE

•

Educational Testing Service (ETS) is strongly committed to the principles of openness in testing, public accountability, quality and fairness. In October 1981, the ETS Board of Trustees adopted and publicly announced as corporate policy the *ETS Standards for Quality and Fairness*. At the same time, the Trustees directed ETS management to maintain a program for monitoring adherence to the Standards and authorized the appointment of a Visiting Committee of persons outside ETS who were to annually review and report to the Trustees on ETS's adherence to the Standards. These actions by the Trustees are tangible evidence of ETS's commitment as a private, nonprofit educational organization to public accountability and to publicly declared standards by which the organization is prepared to be judged. ETS believes that the Standards contribute significantly to the quality and utility of its programs for those institutions and individuals ETS serves.

Compliance with these Standards is taken seriously at ETS. The Standards are applied to all ETS-administered programs. Adherence to the Standards is regularly assessed through a carefully structured audit process and subsequent management review. Every three years, the policies and practices of each program are reviewed by teams of ETS and outside professionals that are asked to report to senior management any instance in which the program does not meet the intent of the procedural guidelines. The audit is a rigorous process. Management then evaluates every recommendation made by the audit teams and decides what action, if any, should be taken to address the teams' findings. It is only at this stage — with the full attention of senior management — that considerations of such factors as cost and technical feasibility are taken into account in judging how to conform to the Standards.

The ETS Standards and their implementation are important matters to the ETS Trustees. To ensure that the Standards are interpreted and applied according to the spirit and purpose intended, the Trustees established a Visiting Committee of persons outside ETS that is comprised of distinguished educational leaders, experts in testing and representatives of organizations that have been critical of ETS in the past. The Committee meets annually with ETS staff, senior management, and outside auditors and it issues a report directly to the Committee on Public Responsibility of the ETS Board of Trustees in June of each year. The Visiting Committee's report is published by ETS and released in its entirety to the media and to any interested members of the public.

The ETS Standards and our efforts to apply them reflect ETS's determination to hold itself accountable to high standards of performance and to setting high standards for the products and services ETS provides. These efforts have been viewed positively by ETS staff as well as the clients we serve. We take great pleasure in noting the first Visiting Committee's conclusion:

> "We find ETS's effort to maintain and improve the quality and fairness of testing well conducted. We know of no other testing organization with anything comparable. The ETS system of auditing its work is an admirable component of ETS's commitment to public accountability; we applaud ETS's intent to be publicly open about activities in which the public clearly has a legitimate interest, even though ETS is a private organization."

This publication represents a continuation of this commitment. In 1984, the American Educational Research Association, American Psychological Association and National Council on Measurement in Education adopted a comprehensive revision of the *Standards for Educational and Psychological Testing*. The ETS *Standards for Quality and Fairness* had been based on the previous joint standards of these three professional associations. We have revised the ETS Standards in order to stay in the forefront of measurement and the latest thinking of the profession. These revised ETS *Standards for Quality and Fairness*, which were adopted by the ETS Board of Trustees on April 10, 1986, will be reviewed carefully during the next year and will be used in the 1986-87 audit process. Following this trial period, the Standards will be reviewed once again, revised if necessary, and presented to the ETS Trustees for their final approval in 1987.

Gregory R. Anrig

Gregory R. Anrig
President

CONTENTS

Page
Introduction vii
Accountability 1
Confidentiality of Data 3
Product Accuracy and Timeliness 5
Research and Development 7
Tests and Measurement — Technical Quality of Tests 10
  Test Development 11
  Test Administration 14
  Test Score Reliability 16
  Scale Definition 18
  Equating 19
  Score Interpretation 20
  Test Validity 22
Test Use 24
Public Information 26
Glossary 28

INTRODUCTION

The *ETS Standards for Quality and Fairness* are designed to ensure that ETS products and services demonstrably meet explicit criteria in seven areas of basic importance: Accountability, Confidentiality of Data, Product Accuracy and Timeliness, Research and Development, Tests and Measurement, Test Use, and Public Information. The first three sections of the Standards deal with issues that relate to all ETS activities: the responsibilities of ETS to those affected by its activities; the rights to and limitations on access to data collected by ETS; and the control of quality and performance. The remaining sections concern issues relating to ETS's main endeavors: Research and Development, Tests and Measurement, Test Use, and Public Information.

The ETS Standards reflect and adopt the *Standards for Educational and Psychological Testing* jointly issued by the American Educational Research Association (AERA), the American Psychological Association (APA), and the National Council on Measurement in Education (NCME). The ETS Standards, however, are tailored to ETS's particular circumstances and needs. Thus, the Standards may not be useful to organizations whose practices, programs or services differ from those of ETS.

The Standards are comprised of both principles that underlie ETS efforts in each area and policies that govern decision-making and guide the development of more specific goals. The Standards are implemented by ETS management through procedural guidelines that provide more detailed criteria for ETS's diverse programs and services. The Standards are reviewed and revised from time to time to keep abreast of developments in professional practice and research.

Like the *Standards for Educational and Psychological Testing* issued by AERA, APA, and NCME, proper interpretation and implementation of ETS's Standards depends on the seasoned judgment of professional staff. These judgments must be carefully based on research, professional experience, and sound reasoning. The ETS Standards are intended to guide and assist ETS professionals in the flexible and sensitive exercise of professional judgments, not to obviate the need for those judgments. Thus, if adherence to any procedural guideline is infeasible or inappropriate in particular circumstances, or if good professional practice in a particular instance conflicts with the letter of a guideline, then sound practice, consistent with the spirit of the underlying principles and policies, should prevail.

ETS does not have sole responsibility or authority to determine how or whether these Standards will be implemented in activities for which practice or policy is substantially established by a group, individual, or institution other than ETS. These Standards are not intended to establish obligations on the part of ETS to act or intervene in situations where the pertinent responsibility rests primarily outside ETS. However, ETS does encourage and assist groups and institutions in implementing the Standards related to any of their activities that involve ETS products or services.

ETS has committed itself to these Standards and to a continuing program of research and development. As a result, ETS expects to expand the realm of knowledge relevant to its activities and to nurture at ETS and elsewhere the development of thoughtful and sensitive professionals with the skills and sensitivity necessary to apply the principles and the policies embodied in these Standards.

ACCOUNTABILITY

Principle

ETS acknowledges responsibility for the effective stewardship of its resources to the New York Board of Regents which has issued its corporate charter; to the governing boards that sponsor and set policy for programs or services in which ETS products or services are used; to the individuals and committees that advise ETS with respect to appropriate policy for its programs; to the institutions and agencies that use ETS products and services; to persons who take ETS tests (and parents or guardians of minor persons), submit data for use by ETS or for distribution to others, or participate in research and development projects conducted by ETS; and to the professional associations that are concerned with educational and psychological measurement and research.

Policies

A. ETS will furnish appropriate information to those to whom it is responsible so they may make informed, independent judgments as to the effectiveness with which ETS exercises its stewardship.

B. ETS will seek, consider, and, as appropriate, act on the views of those who sponsor, use, or are affected by ETS programs and services.

C. ETS will seek advice on its activities and policies from qualified men and women who are not employed or retained on a regular basis by ETS and who are drawn from appropriate professional disciplines, major philosophies and points of view, different geographic regions, and the major subgroups within the relevant population.

D. ETS will support the activities of professional associations with respect to developing and implementing professional standards or codes, making available the results of current work, and fostering peer review of its activities.

Procedural Guidelines

1. Communicate with sponsors by providing information regularly, by reporting program status in a manner consistent with contractual requirements, and by meeting at least annually so that sponsors can:
   - evaluate ETS services in terms of quality, timeliness, and costs;
   - transmit comments or concerns on which ETS will take prompt and appropriate actions; and
   - express opinions about their program and ETS services directly to senior ETS management.

2

2. Make available technical and other information about products and services so sponsors, agencies, institutions, or potential users may evaluate and comment on them. Include representative materials relevant to intended test user(s). Meet requests for additional information not included in publications within a reasonable time and, if necessary, for a reasonable fee so long as disclosure is consistent with legal, ETS, and sponsor policy and contractual requirements.

3. Provide information to persons who take ETS tests, submit data for use by ETS, or participate in ETS research and development projects so they will know:

   - the sponsor's identity and responsibility;
   - the nature of the activity or project;
   - the probable use of the product, service, or research; and
   - the address to which comments, questions, or criticisms can be submitted.

4. Direct to legal counsel significant proposed new or substantially revised activities for review for compliance with federal statutes, regulations, case law, or state law, as appropriate.

5. Seek advice on program policies and plans, where appropriate, from qualified persons of diverse backgrounds, interests, and experience (e.g., professional disciplines, philosophies, geographic regions, major subgroups, relevant populations of interest) who are not regularly employed by ETS. Inform these individuals about the results of their work within a reasonable period of time.

6. Review publications and other materials to eliminate language or material generally regarded as sexist, racist, or otherwise offensive or inappropriate.

7. Record, process, and report financial information accurately and in accordance with generally accepted accounting principles.

8. Monitor changes in federal statutes, regulations, and case law to assure that ETS activities and operations are in compliance. Compliance with other statutes, regulations, or case law will be evaluated as appropriate.

9. Provide reasonable accommodations with respect to professional responsibilities to permit staff members to attend professional meetings, to contribute to the development of professional standards or codes, to engage in activities of professional interest, and to stay abreast of current developments in related fields.

10. Publish an annual report that provides information about organizational activities and finances.

CONFIDENTIALITY OF DATA

Principle

ETS recognizes the right of individuals and institutions to privacy with regard to information supplied by and about them that may be stored in data or research files held by ETS and the concomitant responsibility to safeguard information in its files from unauthorized disclosure.

Policies

A. ETS will ask individuals to provide information about themselves only if it is potentially useful to those individuals, is necessary to facilitate processing of data, or serves the public interest in improving understanding of human performance. Insofar as possible, individuals should be informed of the purpose for which the information is requested.

B. The right of individuals to privacy regarding information about them that may be stored in the data or research files held by ETS extends both to processed information, such as scores based on test-item responses, and the raw data on which the processed information is based.

C. ETS will protect the confidentiality of data supplied by institutions or agencies about themselves, and so identified, to the extent that such confidentiality does not conflict with ETS's obligations to individuals.

D. ETS will not collect or maintain in its data or research files any critical information that in its judgment cannot be protected adequately from improper disclosure.

E. ETS will encourage the organizations with which it works to adopt policies and procedures that adequately protect the confidentiality of the data transferred by ETS to those organizations.

Procedural Guidelines

1. Inform individuals or institutions to the extent appropriate, before information is collected, of the information's intended use, the conditions surrounding its confidentiality and release, and the length of time the information will be retained.

2. Use identifiable information about an individual or institution only for purposes for which permission has been granted unless additional consent is obtained. Release identifiable information from ETS only with proper consent or prior agreement, or in a manner that assures the confidentiality of the individual or institution.

4

3. Make provision for individuals, on presentation of adequate identification (e.g., signature and data file number), to authorize the disclosure of information about themselves from program data files to any appropriate recipient, provided that disclosure does not violate other ETS or sponsor policies or the privacy of other individuals. If authorization is from a third party by prior agreement with the individual, the individual should be notified when disclosure has taken place.

4. Make provision for individuals or their legal representatives to obtain information about themselves from data files held at ETS. Such release of information must be consistent with sponsor's policies and be allowed only upon the individual's submission of appropriate identifying information and, if necessary, payment of a reasonable fee.

5. Assure that access to electronic, paper, or other forms of confidential data is reasonably safeguarded, especially when such data may be part of a time-sharing network, data bank, or other storage medium involving units outside ETS.

6. Develop clear retention guidelines and procedures for eliminating information from data files in accordance with ETS or sponsor policies or contractual requirements whenever information on individuals is maintained.

7. Provide identifiable data only in a manner consistent with these guidelines unless served with a subpoena or other legal process to provide identifiable information. In that event, inform legal counsel in order to make appropriate efforts to narrow the subpoena or to obtain a court order or other arrangements to minimize the dissemination of that information.

8. Inform every organization with which ETS works of the confidentiality of data transferred by ETS to that organization or collected by it on behalf of ETS so that the organization can protect the confidentiality of such data.

PRODUCT ACCURACY AND TIMELINESS

Principle

The accuracy of ETS's principal products and the timeliness with which they are made available are important parts of the responsibility ETS has undertaken with respect to its sponsors and the diverse public it serves.

Policies

A. ETS will establish standards of accuracy and timeliness for each principal product.

B. ETS will use quality controls that are adequate to assure that its standards of accuracy and timeliness are met.

C. ETS will make realistic delivery commitments and reasonable efforts to meet those commitments.

D. ETS will sacrifice the timeliness of the delivery of information if the desired accuracy of that information is substantially in question.

E. ETS will seek to inform those adversely affected if, subsequent to its release, information has been found not to meet ETS standards of accuracy.

F. ETS will seek to inform those adversely affected if there is a probability that there will be substantial departure from ETS standards of timeliness with respect to a principal product.

Procedural Guidelines

1. Verify and document that all principal products conform to specifications or standards before release by doing as many of the following as appropriate:
   - independently recomputing or visually inspecting an appropriate sample of each product; or
   - assessing the reasonableness of computed information through reviews by technically competent staff; or
   - reviewing and proofing printed material; or
   - assuring adherence to ETS or professional standards through effective peer review.

6 2. Verify and document the accuracy of intermediate products when:

- the information (e.g., answer keys, conversion parameters, algorithms) is critical to the principal product; or
- early detection and correction of errors would facilitate meeting delivery schedules of the principal products.

3. Monitor the accuracy, timeliness, and responsiveness of replies to inquiries through periodic audits and other means.

4. Report to a specified ETS staff member all instances in which a product failed to conform to requirements or to standards of accuracy or timeliness. Resolve discrepant conditions before release of the product unless the cognizant ETS officer has approved release to benefit the majority of product users.

5. Correct any critical information found to be in error after its release and promptly distribute corrected information to those adversely affected by the error.

6. Make provision for individuals to verify scores or other information within a reasonable time. Such requests must be accompanied by appropriate identifying information and, if necessary, a reasonable fee.

7. Establish schedules or other process control methods to assure the timely production of each product or service. If it is likely that a product will be late, take steps (e.g., proper notice to test users) to minimize adverse effect.

RESEARCH AND DEVELOPMENT

Principle

A continuing program of research and development conducted in compliance with professional standards with respect to quality and ethical procedures is necessary to maintain the high quality and social utility of ETS contributions to education and society. This includes basic inquiry to increase understanding of educational processes and human development, public policy, evaluative and applied research in response to the needs of the educational community, the work place and society at large; and research and development to improve ETS products and services. Publication of the results of significant ETS research is of benefit to ETS and the profession because it permits others to use, build upon, or improve ETS work.

Policies

A. ETS will devote appropriate research efforts to the following:

- Improving measurement and education through the discovery and conceptual integration of new principles and understanding. This research will be aimed at extending knowledge of measurement principles and practices, knowledge of the learner and learning processes, of learning environments and educational treatments, of educational institutions, and of the interacting factors that influence human development.
- Improving the technical quality and the utility of ETS products and services. Among the important issues addressed by this research will be problems of test development, reliability and generalizability, equating, validity, and the soundness of test interpretation.
- Responding to the measurement and educational needs of society and creating, improving, and evaluating instruments, systems, and programs of service that meet these needs.
- Special problems faced by subgroups in society involved with test taking. In addition, ETS will encourage analysis by subgroup whenever subgroup interests are pertinent to the research being undertaken.

B. ETS will conduct its research under appropriate review procedures that protect the rights of privacy and confidentiality of human subjects or respondents and of cooperating institutions.

C. ETS will follow procedures to insure that ETS research is of high quality.

D. ETS researchers will adhere to appropriate professional and ethical standards, including those published in ***Ethical Principles in the Conduct of Research with Human Participants***, ***AERA Guidelines for Eliminating Race and Sex Bias in Educational Research and Evaluation***, and *Ethical Standards of Psychologists*.

E. ETS will encourage the dissemination of full accounts of ETS research in the usual professional forums and will provide internal means by which the results of ETS research can be disseminated.

Procedural Guidelines

1. Assure the welfare and the right to confidentiality of human subjects or respondents in each project by following procedures approved by the Committee on Prior Review of Research. Procedures approved by the Committee include obtaining appropriate informed consent; separating participants' names from data and other steps relating to confidentiality; and avoiding any negative consequences of participation.
2. Report the results of research with appropriate care to participants and institutions so that the possibility of misinterpretation and misuse are minimized.
3. Publish or otherwise disseminate the results of research projects unless a justifiable need to restrict dissemination is identified before the research begins.
4. Follow review procedures for research proposals and reports that will assure that research is of high quality. Reviews may include the following considerations:
   - the rationale for the research;
   - the soundness of the design;
   - the thoroughness and care of the data collection and analysis;
   - the reasonableness of the interpretation;
   - the clarity of the exposition; and
   - the soundness of the project planning and management.
5. Provide for a periodic assessment of research and development priorities to assure an adequate balance of resources directed toward:
   - improving knowledge of measurement, occupations, educational processes, and human development;
   - meeting the needs of the educational community and society including subgroups of special interest;
   - improving ETS products and services and the manner in which these products and services are used; and
   - developing new methodologies (including educational, psychometric, and statistical, and technological procedures).

6. Whenever sex, ethnic, racial, or other population groups are pertinent to the research, studies should be designed to allow analyses by subgroup. 9

7. Provide non-ETS researchers with reasonable access to ETS-controlled nonproprietary data so long as the privacy of individuals and organizations and ETS's contractual obligations can be protected. Grant access to data facilitating the re-analysis and critique of published ETS research with the same requirements of confidentiality of individuals and institutions. Encourage other organizations to adopt a similar policy.

10

TESTS AND MEASUREMENT— TECHNICAL QUALITY OF TESTS

This section, which deals with ETS testing activities, is divided into seven subsections that are devoted to test development, test administration, reliability, scale definition, equating, score interpretation, and validity.

Principle

High standards of quality and fairness in constructing, administering, reporting, interpreting, and evaluating ETS tests are central to ETS's capability to function effectively as an educational service and research organization.

Policies

A. ETS will strive to develop tests in which the knowledge, skills, abilities, or personal characteristics measured, procedures followed, and criteria used, will be appropriate to the use for which the test is designed and that will be unbiased with regard to relevant major population subgroups being tested.

B. ETS will establish standards for test-administration processes that minimize variations in test performance due to circumstances or conditions not relevant to the attributes being measured.

C. ETS will establish for its tests a high degree of reliability consistent with the requirements and the purposes of the test.

D. ETS will develop scales for reporting scores in a rational fashion consistent with the requirements and the intended use(s) of the test.

E. ETS will provide equating systems, when appropriate, for the perpetuation of score scales with the highest level of precision practicable.

F. ETS will make available to score recipients data for interpreting scores on ETS tests that foster appropriate use of those scores.

G. Recognizing that test validation is a responsibility of both test users and test developers, ETS will encourage and assist test users in their validation efforts and will make available tests that are designed to meet professionally acceptable standards of validity for the primary purposes of each test.

H. ETS will adhere to appropriate professional standards, such as those published in *Standards for Educational and Psychological Testing* and *Principles for the Validation and Use of Personnel Selection Procedures*.

Procedural Guidelines: Test Development

1. Obtain substantive contributions to the test development process from qualified men and women including persons who are not on the ETS staff and who represent diverse institutions, population subgroups, perspectives, and professional specialties. Document their relevant qualifications and characteristics.

2. Ascertain and document appropriate background information for each test to be developed, including:

   - the test's intended use(s);
   - the population that will take the test, including anticipated major subgroups;
   - the procedures followed for defining the domain to be assessed, a description of the domain, and a description of its relevance to anticipated test use(s).

3. Document information relative to the test being developed, including:

   - the rationale for the item type(s) and test format to be used and whether any background or prior experience factors (e.g., age or cultural background of intended test takers) affected item type or test format selection;
   - the procedures followed for generating test content to represent the domain or to link test and job content;
   - the rationale for the scoring method(s), especially when judgmental processes are used;
   - the item response model, calibration procedures, and the nature of the sample used to estimate parameters when item response theory procedures are used to assemble the test;
   - the rationale and procedures for making branching decisions, for terminating the test, and for scoring the test when adaptive or branching tests are used; and
   - the logical or empirical arguments supporting comparability when multiple methods for presenting items or recording responses (e.g., recording answers in test books, on answer sheets, or with electronic devices) are intended to be used and interpretative guidelines for multiple methods where comparability is not supported.

4. Prepare, with appropriate advice and review, test development specifications for each test that cover the following:

   - Content and Skills — a clear description of what is to be tested, including, where appropriate, critical content to be included in each form, and the relative weight to be given to each part of the domain that is to be measured;
   - Test and Item Format — item types to be used, special requirements regarding directions and sample items or tests;

- Psychometric — the intended level of difficulty of the test; the number of items; requirements regarding the target distribution of item difficulties (when using pretested items); requirements regarding the homogeneity of items within each test or subtest and the correlation between subtests or tests; requirements for equating including the content and statistical specifications for equating items; and the testing time allotted or suggested;
- Sensitivity — requirements for the inclusion of material reflecting the cultural background and contributions of major population subgroups; and
- Scoring — the procedures for scoring, especially when judgmental processes are used.

5. Assure that time requirements are consistent with the test's purpose so that time is not a decisive factor in performance for the large majority of test takers, except for tests designed to measure rate of performance.

6. Have subject matter and test development specialists who are familiar with the specifications and purpose of the test and with its intended population review the test items for accuracy, content appropriateness, suitability of language, difficulty, and the adequacy with which the domain is sampled.

7. Review individual items, the test as a whole, directions, and descriptive materials to assure that:
   - appropriate technical standards such as those contained in ETS item writers' manuals are met;
   - language, symbols, words, phrases, and content that are generally regarded as sexist, racist, negative toward major subgroups, or otherwise potentially offensive, are eliminated except when judged to be necessary for adequate representation of the domain;
   - editorial standards for clarity, accuracy, and consistency are met;
   - clear and complete directions appropriate to the nature of the test and the characteristics of the test takers are provided;
   - typography, format (e.g., test books, screens, tapes) and test-book layout facilitate the task of test takers; and
   - sufficient sample questions are contained in program publications to be representative of test content, item types, and difficulty.

8. Evaluate the performance of individual items by pretesting, pilot testing, reviewing the results of administering similar items to a similar population, or conducting preliminary item analysis before scores are reported.

9. Whenever there are sufficient subgroup members to permit meaningful analysis, study item performance relative to subgroups when consideration of the recommended use(s) of the test and the characteristics of the intended test-taking population, in light of prior research, indicates the need for such studies.

10. Evaluate the performance of each test edition by:
    - carrying out timely and appropriate item and test analyses including analyses for reliability, intercorrelation of sections or parts, and speededness;
    - reviewing the adequacy of fit of item response models to data when item response theory procedures are used to develop, score, or equate the test; and
    - comparing the test's characteristics to its psychometric specifications.
11. Review test content and test specifications periodically to assure their continuing relevance and appropriateness to the domain being tested.
12. Review test editions developed in prior years and their descriptions in publications to assure the continued appropriateness of both content and language for the present test-taking population and the subject matter domain.
13. Analyze major changes in test specifications to assure that they are followed by appropriate consideration of the implications for score comparability and to determine whether test name changes or other cautions to test users about comparisons with earlier tests are necessary.

Procedural Guidelines: Test Administration

1. Provide prospective examinees (or, in some programs, parents or guardians as well) with information in advance of the test administration about the following, as appropriate:

   - the test's intended purpose and what it is designed to measure, typical test items, clear directions for the test and the response method to be used, a description of how scores are derived including formation of composite scores, strategies for taking the test (e.g., guessing and pacing), whether the test contains items not intended to be scored, and the background and experience relevant to test performance;
   - the program procedures and requirements, including test dates, test fees, test center locations, special testing arrangements for handicapped persons or others, test registration, score reporting, score cancellation by examinees, ETS, or the sponsor, and registering complaints; and
   - test administration procedures and requirements, including those related to identification and admission to the test center, materials permitted in or excluded from the testing room, and the consequences of misconduct.

2. Establish test centers that are convenient, nondiscriminating, comfortable, and accessible to all individuals including handicapped persons. Locate test centers in both minority and majority communities to foster accessibility.

3. Advise test center staff of the need to minimize distractions and to make examinees comfortable in the testing situation. Instruct staff to be sensitive to the psychological as well as physical needs of examinees. Direct supervisors to consult with or include on the test center staff, when appropriate, subgroup members, and persons knowledgeable about handicapping conditions.

4. Provide test center staff with a description of the program, the expected candidate population, the duties of staff, and the procedures for:

   - receiving, storing, and distributing test materials to examinees, and returning them to ETS;
   - admitting examinees to the test center, including ID requirements;
   - administering the test to examinees, including handicapped individuals;
   - using appropriate seating plans and assignments and monitoring the testing room to reduce opportunities to obtain scores by questionable means;
   - handling of suspected cheating, misconduct, or emergencies; and
   - reporting irregularities (e.g., disturbances, mistimings, defective test questions or materials, power failures, or misconduct) so that, after review, appropriate action can be taken.

5. Provide test center staff with directions (to be read aloud before the test begins) that cover the recording of answers on answer sheets or via other devices, timing of test sections and breaks, guessing strategies, and the consequences of using unauthorized aids or engaging in other forms of misconduct.
6. Utilize effective and equitable procedures for preventing, identifying and resolving scores obtained by questionable means.
7. Encourage examinees to report any irregularities so that, after review, appropriate action can be taken.
8. Undertake quality control activities (e.g., test center observations, solicitation of suggestions from test administrators and examinees, training of test administrators) as necessary for effective and, when appropriate, secure test administrations.
9. Make tests available to handicapped individuals through special testing arrangements or special test editions, as appropriate.
10. Provide users of locally administered tests with instructions about standardized conditions for administering and scoring the tests.

Procedural Guidelines: Test Score Reliability

1. Provide information to enable test users to judge whether reported test scores, including subscores and combinations of scores, are sufficiently reliable for their intended use(s).

2. Document sources of variation (e.g., test form, content, population of readers, time interval between testing, and other sources of error) over which inferences are intended to be made from reported test scores.

3. Estimate the reliability or consistency of reported test scores by method(s) that are appropriate to the nature and intended use of the test scores and that take into account sources of variance considered significant for test score interpretation.

4. Document the method(s) used to assess the reliability or consistency of the test scores and the rationale for using them, the major sources of variance accounted for in the reliability analysis, and the formula(s) used and/or appropriate references.

5. Document the results of the reliability analysis, including:
   - a reliability coefficient, an overall standard error of measurement, classification consistency, or other equivalent information about the consistency of the test scores;
   - standard errors of measurement or other measures of score consistency for score regions within which decisions about individuals are made on the basis of test scores;
   - the degree of agreement between independent scorings when judgmental processes are used;
   - the adjusted and unadjusted coefficients if reliability estimates are adjusted for restrictions of range; and
   - correlations between short forms of tests, if developed, and the standard form.

6. Document the conditions under which the reliability estimates were obtained, including:
   - the nature of the population involved;
   - the selection procedures for and the appropriateness of the analysis sample, including the number of observations, means, and standard deviations for the analysis sample(s) and any group(s) for which reliability is estimated;
   - the basis for scoring when scores are based on judgments, including selecting and training scorers, and the procedures for allocating papers to scorers and adjudicating discrepancies;
   - the time intervals between testings, the rationale for the time intervals, and the order in which the forms were administered if alternate-form or test-retest methods are used;

- speededness data; and
- correlations of reported subscores within the same test or test battery.

7. Whenever there are sufficient subgroup members to permit meaningful analysis, study the reliability or consistency of reported scores for major subgroups when consideration of the intended use(s) of the test and the characteristics of the intended test-taking population, in light of prior research, indicates the need for such studies.

Procedural Guidelines: Scale Definition

1. Establish scales for reporting scores that are well-constructed throughout their range and in a way that facilitates meaningful score interpretation relative to intended use(s) of the scores.

2. Establish scale values to be reported that do not encourage finer distinctions among test takers than can be supported by the precision of the test.

3. Choose the scale values in a manner that avoids confusion with other scales that are widely used by the same population of score recipients.

4. Document the rationale and the methods used to determine score scales. Account for the following as appropriate:

   - If scores derived from different tests in a program are to be directly compared, take into account in the scaling methods the differences among groups taking the different tests.
   - If the scale is to be normative, consider the probable length of time and the extent to which the normative information will be appropriate and useful for the intended population.
   - If a test or test battery yields multiple scores for an individual and comparisons among scores are encouraged, establish scales in a manner that allows meaningful comparisons among scores (e.g., normatively or against an absolute standard), or provide data to allow such comparisons.
   - If the scale is to be defined with reference to performance standards, classification, or cut scores, document the method and rationale used, and the qualifications of any judges.
   - If a scale is used to report composite scores derived from weighting subscores, clearly state the rationale and the method for weighting the subscores.

5. Avoid reporting raw scores or percentages of questions answered correctly on a test or subtest except under one or more of the following circumstances:

   - only one edition of the test is to be offered;
   - scores on one edition will not be compared with scores on another;
   - raw scores on all editions are comparable; or
   - raw scores are reported in a context that supports the intended interpretation(s).

6. Report item responses for individuals or groups only in a context that supports the intended appropriate interpretation(s).

7. Redefine an established scale only under compelling circumstances. Provide announcements to all score recipients indicating the change and cautioning recipients against comparisons with earlier scores. If the numerical values are to be changed, change them substantially to minimize confusion between the old and the new scale.

Procedural Guidelines: Equating

1. Assure comparability of scores that are derived from different editions of the same test and are used to compare individuals or groups.
2. Document methods used to achieve comparability, including:
   - the rationale for selecting the methods used;
   - the consistency between the assumptions underlying the method and the circumstances under which the method is applied (e.g., when test editions are equated using common items, make the directions, context, speededness, item placement, and other aspects of the test nearly the same as possible for all examinees; when anchor scores are based on a test that is not representative of the tests being equated, make sure the groups of examinees used for equating are equivalent; or when item response models are used, make sure that information is presented on the adequacy of fit of the model to the data);
   - the procedure for linking adequately all editions of the test for which scores should be comparable; and
   - the plans for specially designed studies to collect data to achieve comparability if only a limited number of editions are offered to institutional or other users who will administer and score the tests.
3. Document the results of the equating experiment including:
   - the nature of the population involved;
   - a description of the analysis sample(s), including the number of observations, means, and standard deviations;
   - the time intervals between testings; and
   - other statistics appropriate to the method used (e.g., correlation between the anchor test, if used, and the total test).
4. Periodically assess the results of methods used to achieve comparability of scores and evaluate the stability of the score scale.

Procedural Guidelines: Score Interpretation

1. Provide score interpretation information for all score recipients in terms that facilitate appropriate interpretations. Provide information that is appropriate for each category of score recipient (e.g., examinee, teacher, college, agency, or media) and that minimizes the possibility of misinterpretation of individual scores as well as group results.

2. Provide each category of score recipients with appropriate information that:

   - concerns the intended use(s) of the test and what it is designed to measure;
   - recommends only those score interpretations for which supporting information is available;
   - describes scale properties that affect score interpretation and use;
   - explains the variability of and limitations on the accuracy of test scores (e.g., standard error of measurement, classification errors), and encourages recipients to take such information into account in making decisions based on scores;
   - supports assessments based on individual items or clusters of items whenever such uses are suggested; and
   - gives the minimum score(s) required to pass the test when results are reported as pass fail and examinees have failed the test.

3. Provide score recipients with an appropriate frame of reference for evaluating the performance represented by test scores through information based on norms studies, carefully selected and defined program statistics, or logical analysis. When statistical information is included, the information should be adequately labeled and the nature of the group(s) on which the information was based should be clearly identified.

4. Document the method(s) (e.g., norms studies, derivation of program statistics, cut-score studies) used to develop score interpretation information. Provide the following types of information, as appropriate:

   - the characteristics of the scale and procedures used to maintain it;
   - the method of selecting participants on which data are based, including information about representation of relevant major subgroups within the defined population;
   - the participation rate of categories of individuals or institutions and their characteristics such as the age, sex, or subgroup composition of the group, weighting systems or other adjustments made to form the norming sample, and whether or not the participants were self-selected;
   - the period in which the data were collected;
   - appropriate group statistics whenever tests are intended to be used to make assessments of such groups (e.g., classrooms) rather than individuals;

- methods and rationale for aggregating test results or developing composite scores;
- estimates of sampling error and possible effects of nonparticipation;
- comparisons with relevant data on variables from other sources when possible; and
- evidence supporting the cut scores or configural scoring rules when different score interpretations are automatically provided for examinees scoring at different points on the scale.

5. Revise norms or other score interpretation information at sufficiently frequent intervals to assure its continued appropriateness as a frame of reference for evaluation of performance represented by test scores.
6. Compile descriptive statistics periodically from samples or from the entire population to monitor the participation and performance of major subgroups.
7. Provide score recipients with information as appropriate to assist them in using scores in conjunction with other information, setting cut scores, interpreting scores for major subgroups, conducting local norms studies, and developing local interpretive materials.
8. Avoid developing interpretive information for subgroups unless sufficient data are available on each subgroup to make the information meaningful, the information can be accompanied with a carefully described rationale (e.g., guidance purposes) for using it, and the information can be presented in a way that discourages incorrect interpretation and use.
9. Caution score recipients, when appropriate, that:
    - scores for different tests offered by a program may not be comparable even though the scores are reported on similar scales;
    - inferences that have not been adequately validated (e.g., ones based on foreign language translations, untimed tests for handicapped persons, experimental tests) should be made with care;
    - scores may no longer be comparable if test content or specifications have changed sufficiently;
    - scores earned in previous years may become of limited value due to changes in the individual or the meaning of test scores over time; and
    - decisions based on the differences between test scores for an individual (e.g., aptitude and achievement) should take into account the overlap between the constructs and the reliability of the score difference.

Procedural Guidelines: Test Validity

1. Provide evidence relating to the intended use(s) of the test scores. Assure that tests are validated by procedures that are most appropriate to the intended use(s) of the test scores.
   - Content-related evidence generally is based on a description of how the test and test items were derived from and are related to the areas of interest.
   - Criterion-related evidence generally is based on statistical relationships between test scores and as many distinct performance variables as necessary to evaluate the test score's effectiveness.
   - Construct-related evidence generally is based on the logical and empirical analysis of processes underlying performance on the test, the relationship between test scores and other pertinent variables.
2. Describe how the validity evidence provided is appropriate to the intended use(s) of the test.
3. Document the validation procedures used and the results of the analyses performed. Address the following points, as appropriate:
   - the number and qualifications of any experts who made judgments, and procedures used to arrive at judgments pertinent to the validation effort;
   - the materials surveyed, and the rationale and procedures for defining test content;
   - for tests designed to sample job functions, the link between job tasks and test content and, when specified, the link between job tasks and the knowledge, skills, and abilities being tested;
   - the rationale and procedures for determining criterion relevance, the selection procedures for and the composition of the validation sample, the relationship between predictors and criteria, and factors that affect the relationship, including technical quality of the criteria (e.g., their reliability, the elapsed time between test administration and criterion data collection, and rules for combining criteria if several criteria are combined); and
   - when quantitative evidence is reported, information relative to its interpretation such as associated standard errors of the estimate, adequacy of the sample, possible restriction of range of scores on the variables, unadjusted coefficients (when statistical adjustments are made), the need for cross validation, and other contextual factors.
4. Base validity evidence in a particular situation (e.g., institution, department, or job study) on data from other situations only when it can be established that the particular situation is from the same population of situations. Include in documentation information about the similarity of the groups tested, the curricula, the job tasks, or other appropriate criterion variables.

5. Undertake new validity studies whenever the test, mode of administration, the characteristics of the intended test-taking population, or the performance domain sampled is changed substantially. 23

6. Whenever there are sufficient subgroup members to permit meaningful analyses, investigate validity for major subgroups when consideration of the intended use(s) of the test scores and the characteristics of the intended test-taking population in light of prior research indicates the need for such investigation.

7. Establish test names that imply no more than the validity evidence justifies.

8. Provide information to users to help them plan, conduct, and interpret validity studies.

TEST USE

Principle

Proper and fair use of ETS tests is essential to the social utility and professional acceptance of ETS work.

Policies

A. ETS will set forth clearly to all score recipients the principles of proper use of tests and interpretation of test results.

B. ETS will establish procedures by which fair and appropriate test use can be promoted and misuse can be discouraged or eliminated.

Procedural Guidelines

1. Provide score recipients (e.g., examinees, teachers, colleges, agencies, or the media) with adequate descriptions of intended test use(s), caution them about making interpretations not supported by validity evidence, and warn them against reasonably anticipated misuses.
2. Encourage test users to put test scores in an appropriate perspective (e.g., augment test scores with other relevant information about the examinee, provide multiple opportunities to retest or to demonstrate relevant skills by other means).
3. Provide users with opportunities for consultation about test use and with information about reliability, validity, test content, test difficulty, and representative research.
4. Advise users that when using test scores differently for members of different subgroups (e.g., separate sex norms or using racial data in regression equations), such uses should be carefully and rationally supported.
5. Advise users that whenever individuals are assigned to groups on the basis of test scores, users should undertake periodic examinations of:
    - pass-fail or cut-score policies;
    - the rationale and methods for making assignments;
    - the performance of individuals within their respective groups, where feasible, including the collection of empirical evidence to support the assignments;
    - the continued appropriateness of assignment criteria; and
    - classification rates across major subgroups.

6. Investigate complaints or allegations of improper score use. When a misuse is verified, advise the sponsor and the user and seek voluntary correction. If efforts to achieve voluntary correction are not successful, consult with the sponsor to determine whether to continue services to the misuser. Maintain records of complaints and their disposition.
7. Assure the accuracy of any ETS-produced promotional material concerning tests and their intended uses.

PUBLIC INFORMATION

Principle

ETS is dedicated to promoting public understanding of testing, measurement, and related educational issues by providing programs of public information, research, and advisory and instructional activities.

Policies

A. ETS will promote understanding of the purposes and procedures of testing and the proper uses of test information among examinees, test users, and the general public; ETS will encourage sponsors to undertake similar efforts.

B. ETS will adhere to high professional and ethical standards in both the promotion and the use of its products and services and in the dissemination of information to examinees, test users, and the general public. ETS will encourage sponsors and other organizations to do so.

C. ETS will provide instruction and technical assistance in testing, measurement, evaluation, and related areas.

D. ETS will disseminate the results of research on testing, measurement, and other related educational issues and will make ETS-controlled nonproprietary data available to other researchers; further, ETS will encourage other organizations to do the same.

E. ETS will respond promptly and appropriately to requests for advice and technical assistance related to programs and services offered by ETS, to purposes and procedures for testing, to uses and misuses of test information, and to complaints about its services.

F. ETS will collect reference materials relating to tests, measurement, evaluation, and related research, and will make its collections available to professional groups, organizations, and interested individuals.

Procedural Guidelines

1. Develop and disseminate publications and other materials to promote proper test use, discourage misuse, and improve public understanding of testing, measurement, and related educational issues directly and in collaboration with sponsors.

2. Convene periodically groups of test users, measurement specialists, representatives of professional groups, and other interested parties to examine ETS procedures and recommend improvements in them.

3. Provide accurate and appropriate information when marketing ETS products and services. 27

4. Provide advice and technical assistance on tests and measurement for test sponsors, users, and other interested groups.

5. Offer conferences, seminars, workshops, and other forms of training or instruction in testing, measurement, and other relevant areas of interest, acting independently or in cooperation with other institutions or professional groups.

28

GLOSSARY OF TERMS

Absolute Standard: A cutscore or performance standard that is established without reference to the score distribution of the people for whom the standard will be operational. For example, a passing score set at 80 percent of the questions correct without basing the decision on how many people will score above or below that point is an absolute standard. See *Cutscore, Performance Standard.* Compare *Relative Standard.*

Accuracy: The extent to which a principal product conforms to its specifications.

Achievement Test: A test that measures a particular body of knowledge or set of skills and that is ordinarily used to assess a person's level of performance after the person has participated in some learning experience, the outcome of which the test is intended to measure. Compare *Aptitude Test.*

Adaptive Test: A test administered such that the next item to be administered to a person depends on the person's response to a previous item or set of items.

Adjusted Coefficient: A statistic that has been revised to estimate its value under conditions other than those in the sample on which it has been calculated. For example, a correlation coefficient may be adjusted to account for restriction of range. See *Restriction of Range.*

Alternate Form: An edition of a test that is written to meet the same specifications and is comparable in most respects to another edition of the test except that some or all of the questions are different. An alternate form may or may not be a parallel form. Compare *Parallel Form.* See *Test Specifications.*

Alternate Form Reliability: An estimate of reliability based on the correlation between alternate forms of a test administered to the same group of people. See *Alternate Form, Reliability.* Compare *Internal Consistency Reliability, Test-Retest Reliability.*

Analysis Sample: The group of people on whose performance a statistic or set of statistics has been calculated.

Anchor Test: A usually relatively short test administered with two or more forms of a test for the purpose of equating those forms. See *Common Items, Equating.*

Answer Key: A listing of the correct responses to a set of test questions.

Aptitude Test: A test that is usually not closely related to a specific curriculum and which is used primarily to predict future performance. Compare *Achievement Test.* Note that the distinction between aptitude tests and achievement tests is not strong and depends more on differences in test use than on differences in test content.

Attributes: Qualities or characteristics of a person, such as command of a body of knowledge, ability to perform certain skills, or interest in performing a particular type of task. 29

Branching Test: See *Adaptive Test.*

Classification Error: (1) The proportion of inconsistent categorizations of examinees that would be made on repeated administrations of the same test or of a test and an alternate form, assuming no changes in the examinees' true performance levels. (2) The assignment of an examinee to the wrong category, such as passing a person who lacks minimal competence and should fail.

Classification Rates: The proportions of examinees placed in various categories, such as pass-fail, on the basis of test scores.

Client: (See Sponsor)

Committee on Prior Review: An ETS institutional review board that reviews proposed and ongoing research to ensure adequate protection of human subjects.

Common Items: A set of test questions that remain the same in two or more forms of a test for purposes of equating. The common items may be dispersed among the items in the forms to be equated or kept together as an anchor test. Compare *Anchor Test.* See *Equating.*

Comparable Scores: Scores that are put on the same scale so that they have the same meaning in terms of relative ranking within a defined group of people but that cannot necessarily be used interchangeably. For example, percentile rank scores on a reading test and on a math test are comparable scores if the percentile ranks have been based on the same norm group for both tests. Compare *Equivalent Scores.*

Composite Score: A score that is the combination of two or more scores by some specified formula.

Configural Rule: A specified procedure for interpreting the pattern of a person's scores on two or more tests or subtests.

Consent: Permission granted by an individual or that individual's parent or guardian for the use or release of data held by ETS; such permission granted upon receipt of a reasonable explanation of the purpose of the use or release and a reasonable explanation of the manner in which the results will be reported.

Construct: A theoretical concept developed to explain a group of related behaviors. Examples of constructs are "intelligence," "creativity," "self concept," "anxiety."

Conversion Parameters: Quantitative rules for expressing scores on one test form in terms of scores on an alternate form. See *Alternate Form, Equating.*

Criterion: (1) That which is predicted by a test, such as college grade-point average or job-performance rating. (2) The score with which responses to a test item are correlated.

Criterion Relevance: The extent to which the measure used in assessing a test's predictive validity is related to the test's intended purpose.

Critical Content: Knowledges, skills, or abilities that must be measured in a test because of their importance.

Critical Information: Information that will be used to draw important inferences (a) about the sponsor, ETS-appointed external committees, institutional or agency user, examinee, subject or respondent, or (b) by the sponsor, institutional or agency user, examinee, subject or respondent and which, if incorrect, could be harmful.

Cross Validation: The application of scoring weights or prediction equations derived from one sample to a different sample to allow estimation of the extent to which chance factors determined the weights or equations or inflated the validity estimated in the analysis sample.

Cutscore: A point on a score scale at or above which examinees are classified in one way and below which they are classified in a different way. For example, if a cutscore is set at 60, then people who score 60 and above may be classified as "passing" and people who score 59 and below classified as "failing."

Domain: A defined body of knowledge, skills, abilities, attitudes, interests or other characteristics.

Equating: A statistical process used to convert scores on two or more alternate forms of a test to a common scale such that the scores may be used interchangeably. See *Anchor Test, Common Items, Conversion Parameter.*

Equivalent Scores: Test scores that can be used interchangeably. Compare *Comparable Scores.*

ETS Board of Trustees: The ETS Board of Trustees is the governing body of ETS. There are 17 trustees. Sixteen are elected for four-year terms. New members of the Board are elected by current trustees. The President of ETS is an *ex officio* member.

ETS-Held Program Data Files: Information about individuals and institutions held by ETS and derived from ETS-provided services of collection, processing, storage, retrieval and dissemination.

ETS-Held Research Files: Information held by ETS and generated through ETS-conducted research.

Examinee: An individual who takes a test, developed and/or administered by ETS.

Formula Score: Raw score on a multiple choice test after a correction for guessing has been applied, usually the number right minus a fraction of the number wrong. See *Raw Score.*

Handicapping Conditions: (1) A visual, auditory, other physical or learning disability such that a test administered under standardized conditions would result in a score that significantly underestimates the person's true ability. (2) A disability which limits a person's access to a testing site. See *Standardized Conditions.*

Institutional or Agency User: An organizational recipient of ETS-processed or produced information.

Intermediate Product: Materials that are not released externally, but that are necessary to the production of the principal product.

Internal Consistency Reliability: An estimate of reliability based on the extent to which the items in a test tend to measure the same attribute in the same way. See *Reliability*. Compare *Alternate Form Reliability*, *Test-Retest Reliability*.

Item: A test question.

Item Analysis: A statistical description of how an item performed within a particular test when administered to a particular sample of people. Data often provided are the difficulty of the question, the number of people choosing each of the options, and the correlation of the item with some criterion.

Item Response: (1) A person's answer to a question. (2) The answer to a question coded into categories such as right, wrong, or omit.

Item Response Theory: A set of propositions relating people's performance on test questions to certain characteristics of the people and certain characteristics of the items by means of mathematical models. It is based on the assumption that the probability of a correct response by a person to an item can be calculated from the examinee's estimated ability and certain statistical characteristics of the item.

Item Type: The observable format of a test question. At a very general level "item type" may refer, for example, to multiple choice or free response questions. At a finer level of distinction, "item type" may refer, for example, to synonym questions or antonym questions.

Local norms: A distribution of scores and related statistics within an institution or closely related group of institutions (such as the schools in one district) used to give additional meaning to test scores by serving as a basis for comparison.

Locally Administered Test: A test that is given by an institution at a time of the institution's own choosing.

Normative Scale: A way of expressing a score's relative standing in the distribution of scores of some specified group.

Parallel Forms: Alternate forms of a test that yield nearly identical means and standard deviations of scores as well as nearly identical correlations between scores and other variables. See *Alternate Forms*.

Parameter: (1) The value of some variable for a population as opposed to an estimate of the value based on a sample drawn from the population. (2) In item response theory, one of the characteristics of an item such as its difficulty.

Part Score: A score derived from a subset of the items in a test. Synonym of *Subscore*.

Performance Standard: A cutscore or a defined level of performance on some task. For example, "Run 100 yards in 12 seconds or less." See *Cutscore*.

Pilot Testing: Small scale try-out of test questions or a test form often involving observation of and interviews with examinees.

Precision: The width of the interval within which a value can be estimated to lie with a given probability. The higher the precision, the smaller the interval required to include the value at any given probability.

Principal Product: ETS-produced or processed materials (e.g., annual reports, performance data, score reports and admissions tickets) that are released or transmitted to a sponsor, ETS appointed external committee, institutional or agency user, examinee, subject or respondent, pursuant to a contract or published commitment.

Principles For The Validation And Use Of Personnel Selection Procedures. Division of Industrial-Organizational Psychology, American Psychological Association, Berkeley, CA: The Industrial-Organizational Psychologist, 1980.

Program Statistics: Data that are based on the groups of people that happen to take the tests offered by a particular testing program. Program statistics are not equivalent to data derived from carefully selected samples of defined populations such as those used to construct national norms.

Raw Score: (1) The number of items answered correctly on a test with no adjustment. (2) In some usages, the formula score is also called a raw score. See *Formula Score.*

Regression Equation: A formula used to estimate the value of a variable given the value of one or more observed variables. For example, estimating college grade point average given high school grade point average and SAT scores.

Relative Standard: A cutscore or performance standard that is established with reference to the score distribution of the people for whom the standard will be operational. For example, a cutscore set to pass 60 percent of the people is a relative standard. See *Cutscore*. Compare *Absolute Standard.*

Reliability: An indicator of the extent to which test scores will be consistent across different conditions of administration and/or administration of alternate forms of the test. See *Alternate Form Reliability, Test-Retest Reliability.*

Respondent: An individual who provides data to a research project in a manner and for a purpose different from either examinees or subjects.

Response Method: The procedure used by an examinee to indicate an answer to a question such as a mark on an answer sheet, a handwritten essay, or an entry in an electronic storage medium.

Restriction of Range: A case in which the variance of scores in an analysis sample is lower than the variance of scores in the population from which the sample was selected. See *Analysis Sample, Variance.*

Sampling Error: The difference between a statistic derived from a particular sample and its value in the population from which the sample was drawn. See *Parameter (1).*

Score: A quantitative or categorical value (such as "pass" or "fail") assigned to an examinee as the result of some measurement procedure.

Score Recipient: A person or institution obtaining the scores of individual examinees or summary data for groups of examinees.

Score Scale: The set of numbers within which scores are reported for a particular test or testing program, often, but not necessarily, having a specified mean and standard deviation for some defined reference group.

Special Testing Arrangement: A test administered under non-standardized conditions in which modifications have been made to meet the needs of examinees who require the modifications for appropriate assessment such as providing audio-taped versions of tests for visually-impaired people. See *Standardized Conditions.*

Speededness: The extent to which peoples' scores are affected by how quickly they respond to items on a test. One indicator of speededness is the percent of test takers who answer all of the items in the test.

Sponsor: Educational, professional or occupational associations, federal, state or local agencies, public or private foundations which contract with ETS for its services. This category includes their governing boards, membership and appointed committees or staff.

Standard Deviation: A statistic characterizing the magnitude of the differences among a set of measurements. Specifically it is the square root of the average squared difference between each measurement and the mean of the measurements. See *Variance.* The standard deviation is the square root of the variance.

Standard Error of Estimate: A statistic that indicates the standard deviation of differences between actual and estimated measures. It is an indicator of the accuracy of the estimate. See *Standard Deviation.*

Standard Error of Measurement: A statistic that indicates the standard deviation of the differences between observed scores and their corresponding true score. It has also been described as the standard deviation of scores for a person taking a large number of parallel forms of a test, assuming no changes in the person's true ability. See *True Score, Standard Deviation.*

Standardized Conditions: The administration of a test in the same manner to all examinees to allow fair comparison of their scores. Factors such as timing, directions, use of aids such as calculators and dictionaries are controlled to be constant for all examinees.

Standards for Educational and Psychological Tests. American Psychological Association (APA), American Educational Research Association (AERA), and National Council on Measurement in Education (NCME). Washington, D.C.: APA, 1985.

Subgroup: A part of the larger population which is definable according to various criteria as appropriate, (e.g., by sex, race or ethnic origin, training or formal preparation, geographic location, income level, handicap and/or age).

Subject: An individual who participates in an ETS laboratory or experimental research project.

Subscore: A score derived from a subset of the items in a test. Synonymous with *part score*.

Subtest: A subset of the items in a test upon which a subscore or part score is based.

Test Analysis: A description of the statistical characteristics of a test following administration, including but not limited to distributions of item difficulty and discrimination indices, score distributions, mean and standard deviation of scores, reliability, standard error of measurement, and indices of speededness.

Test Battery: (1) A collection of measures designed to allow the comparison of scores across measures for an individual. (2) Loosely speaking, a collection of tests often administered together.

Test Form: A unique edition of a test consisting of all of the identical copies of a test. Compare ***Alternate Form, Parallel Form.***

Test Format: The physical layout of a test including the spacing of items on a page, type size, positioning of item response options, etc.

Testing Program: A set of arrangements under which examinees are scheduled to take a test under standardized conditions, the tests are supplied with instructions for giving and taking them, and arrangements are made for scoring the tests, reporting the scores, and providing interpretive information as part of a comprehensive ongoing service. A program is characterized by its continuing character and by the inclusiveness of the services provided.

Test-Retest Reliability: An estimate of reliability based on the correlation between scores on two administrations of the same test to the same group of people. See ***Reliability***. Compare ***Alternate-Form Reliability***.

Test Specifications: Detailed documentation of the intended characteristics of a test including but not limited to the content and skills to be measured, the number and type of items, the level of difficulty, the timing, and the layout.

Test-Taking Population (Intended): The people for whom a test has been designed to be most appropriate. The actual test taking population may differ in some instances from the intended population.

Timeliness: The degree to which a principal product is released or delivered to its recipient within a predefined schedule.

True Score: The hypothetical average score of an examinee calculated from an infinite number of administrations of equivalent test forms assuming no learning, forgetting, or fatigue on the part of the examinee. It is the score that an examinee would obtain if the test were perfectly reliable and the standard error of measurement were zero. See ***Reliability, Standard Error of Measurement.***

Validity: The extent to which inferences made on the basis of test scores are appropriate and justified by evidence.

Variance: A statistic characterizing the magnitude of the differences among a set of measurements. Specifically it is the average squared difference between each measurement and the mean of the measurements. 35

Weighting System: (1) A formula giving the relative contribution (expressed as a multiplier) of part scores to a composite score. See *Composite Score.* (2) The relative contribution assigned to certain sample data to better represent a target population.

The ETS Sensitivity Review Process: Guidelines and Procedures

ACKNOWLEDGMENT

This document is a revision of the original *ETS Test Sensitivity Review Process* that was developed in 1980 by Ronald V. Hunter and Carole D. Slaughter.

Substantial contributions have been made by other writers of earlier documents dealing with the issue of sensitivity. Many of these pioneering efforts, such as the *ETS Guidelines for Testing Minorities*, the *ETS Guidelines for Sex Fairness in Tests and Testing Programs*, and the *Guidelines for Avoiding Sexist Language*,* provided much of the creative thought and detail contained within this document.

Finally, many ETS staff members have taken the time to review drafts of this document. in so doing they have provided a wealth of helpful suggestions and productive insights on this complex issue.

---

*From McGraw-Hill, *Guidelines for Equal Treatment of the Sexes*, 1974. Used with the permission of McGraw-Hill Book Company. Recently reissued in *Guidelines for Bias-Free Publishing*.

Copyright © 1986 by Educational Testing Service
All rights reserved

Educational Testing Service, *ETS*, and are registered
trademarks of Educational Testing Service.

Educational Testing Service is an equal opportunity
affirmative action employer.

Table of Contents

Introduction 3

Process Overview 3
  Reviewers 3
  Sensitivity Review for Tests 3
  Sensitivity Review for Other Publications 4

Procedures for Sensitivity Reviews of Tests 4
  Preliminary Review 4
  Final Review 4
  Arbitration Process 6

Procedures for Sensitivity Reviews of Publications 7
  Mandatory Review 7
  Arbitration Process for Publications 7

Evaluation Guidelines 8
  Definitions 8

Evaluation Requirements 11
  Cognitive/Affective 11
  Controversial Material 14
  Examinee Perspective 14
  Balance 14
  Stereotyping 15
  Caution Words and Phrases 15
  Special Review Criteria for Women's Concerns 15
  Special Review Criteria for References to People with Disabilities 15
  Underlying Assumptions 16
  Context Considerations 16
  Elitism, Ethnocentricity, and Related Problems 16

Appendix A: Guidelines for Recognition of Unacceptable Stereotypes 17

Appendix B: Caution Words and Phrases 19

Appendix C: Special Review Criteria for Women's Concerns 22

Appendix D: Special Review Criteria for References to People with Disabilities 25

Appendix E: Sample Forms 27

Appendix F: Sensitivity-Related Sections of ETS Official Documents 31

ETS SENSITIVITY REVIEW PROCESS: GUIDELINES AND PROCEDURES

INTRODUCTION

Educational Testing Service is committed to ensuring that its tests and publications acknowledge the multicultural and multiethnic nature of our society and reflect a thoughtful and fair consideration of the very broad character of ETS's clientele. As part of the effort to attain this goal, ETS has stated in its *Standards for Quality and Fairness* that individual test questions, tests as a whole, and descriptive materials must not contain language, symbols, words, phrases, and examples that are generally regarded as sexist, racist, or otherwise potentially offensive, inappropriate, or negative toward any group.[1]

This document is the basic guide to the process through which these standards are met. It identifies the sensitivity criteria used in the reviews and details all review procedures. Although most of the criteria are general ones that can and should be applied to any population group, experience has shown that a special effort must be made to evaluate material from the perspectives of Asian/Pacific Island Americans, Black Americans, Hispanic Americans, individuals with disabilities, Native Americans/American Indians, and women. This publication, therefore, specifically addresses areas of special concern to these six groups.

PROCESS OVERVIEW

Reviewers

The reviewers for sensitivity evaluations are trained in two-day workshops that cover all issues; in addition, there are one-day refresher workshops for periodic review of sensitivity issues. Trained staff members from test development and test editing areas represent the general disciplines of the humanities, the social sciences, the sciences, and vocational education. Trained editorial staff also serve as sensitivity reviewers for nontest publications. While women and minority staff members are represented among the reviewers, any professional volunteer can be trained to perform sensitivity reviews. Before formally reviewing test material or other ETS publications, all reviewers receive training in the ETS sensitivity guidelines and the process in order to ensure that they understand the review criteria and are able to apply them consistently.

Sensitivity Review for Tests

The test sensitivity review process has three major components: an optional preliminary review, a mandatory final review, and an arbitration process. Every pretest and final form (scored test) must have a sensitivity review, and every test more than five years old must have one before reprinting.

Preliminary Review (optional): Any staff member in the process of assembling a test may request a preliminary review to screen questions, reading passages, and other such materials for possible problems and deficiencies. The reviewer's recommendations are not binding at this stage. However, this review may reveal problems at a point early in the test development process when modifications can be made more easily.

Final Review: The mandatory final review takes place at the time of the editing process. After editing, substantive changes are not normally made in a test. This final sensitivity review must be conducted, even if the test received a preliminary review. If possible, the preliminary and final reviews should be performed by the same person.

---

[1] See Appendix F

Arbitration Process: If the person who assembled the test and the sensitivity reviewer cannot agree on how to resolve the issues raised by the reviewer, the two parties meet with the sensitivity review coordinator from the sensitivity reviewer's area. The coordinator acts as a mediator. If the issue still is not resolved, the three parties meet with the test development director from the test assembler's area to discuss a possible resolution. If mediation is unsuccessful at this stage, the material in question goes to arbitration.

An arbitration panel consists of three staff members not in test development divisions. These arbiters receive the same training in the ETS sensitivity guidelines as do the reviewers. Arbiters may not serve on a decision-making panel involving a program for which they work.

After examining the disputed material, the panel decides whether it violates the guidelines. As part of this process, the panel may choose to review the entire test and to address any issues it may find in addition to those submitted for arbitration.

The decisions of the arbitration panel are final and binding.

Sensitivity Review for Other Publications

Sensitivity reviews for nontest publications are conducted as part of the normal editorial review process. Ordinarily, sensitivity issues are resolved between the reviewer and the author. In case of disagreement, the dispute is resolved through the same arbitration process used for test material.

PROCEDURES FOR SENSITIVITY REVIEWS OF TESTS

Preliminary Review

During the optional preliminary review, test items, reading passages, and other such materials can be screened to detect potential problems. The preliminary review is performed at the request of the test assembler, who provides a reviewer with the test work folder, which contains several documents, including the following:

1. A copy of the test specifications
2. The test items (usually unassembled)
3. Any other relevant material
4. The test sensitivity review report form

The sensitivity reviewer returns the work folder and report form (see Appendix E) with comments and recommendations to the test assembler within 48 hours. Time is charged to the project/job for the test.

Although the reviewer's recommendations are not binding, failure to modify the test material might result in similar recommendations during the final review. As the need to modify a test to any significant degree during the editing process (final mandatory review) can cause delays in the overall test development process, test assemblers are encouraged to use the preliminary review for any material that might raise sensitivity issues.

Final Review

The mandatory final review takes place during the test-editing process.[2] If the test has received a preliminary review, another sensitivity review must be performed; it will be acceptable to redate the preliminary form (*if the mandatory review reveals no problems*) to indicate that the mandatory review has been performed. It is recommended that the preliminary and final reviews be performed by the same person. The test assembler may request that a subject specialist review the test when content is critical. The steps to be followed for the final review are:

1. The test assembler fills out the top portion of the front page of the test sensitivity review report form. It is important to indicate at this time the exact nature of both the final form requirements and pretest requirements for multicultural material in the test.

---

[2] Test editors perform sensitivity reviews as part of the editing process for some mathematics and science tests approved for such reviews by the test development directors. The editor signs the test assembler's control sheet indicating that a sensitivity review has been performed.

2. The test assembler submits the entire test work folder to the sensitivity review router in his or her division for assignment to a sensitivity reviewer.
3. The sensitivity review router logs the work folder and assigns it to a reviewer. The router may give the test to a reviewer from another division.
4. The sensitivity reviewer evaluates the test in accordance with the *Guidelines* to determine conformity.[3]
5. The sensitivity reviewer completes the test sensitivity review report form, by which sensitivity comments and recommendations are documented, and returns it to the router along with the work folder within 48 hours. If no recommendations are made, the sensitivity reviewer indicates acceptance on the test assembler's control sheet and the test sensitivity review form and returns them to the router along with the work folder. Time is charged to the project/job for the test.
6. The test assembler discusses the report with the sensitivity reviewer as necessary. If the sensitivity reviewer has made no recommendations, the assembler signs and dates the report form and files it in the work folder. The test or test section is then sent through the usual test production cycle. If the sensitivity reviewer has made recommendations, the test assembler provides a written response to each issue, outlining planned action and, where appropriate, a rationale, and returns these responses to the sensitivity reviewer along with the work folder.
7. The sensitivity reviewer indicates concurrence or nonconcurrence with the test assembler's responses and returns the form and work folder within 48 hours of receipt to the test assembler.

   If the sensitivity reviewer is satisfied with all of the responses, the report form is signed and dated, the control sheet is signed, and all documents are returned to the test assembler along with the work folder. The time is charged to the project/job for the test.

   A sensitivity reviewer who disagrees with the test assembler's planned actions will meet first with the test assembler and the sensitivity review coordinator from the test assembler's area. If no resolution occurs, the assembler, the reviewer, and the sensitivity review area coordinator from the test assembler's area meet with the test development director of the test assembler's area to attempt to resolve the issue(s).

   The test development director serves as a mediator and attempts to resolve the issues to the mutual satisfaction of both the sensitivity reviewer and the test assembler.[4] If the problem is resolved at this time, one of two processes takes place:

   a. The sensitivity reviewer indicates concurrence with the test assembler's rationale, and both the sensitivity reviewer and test assembler sign and date the report form, indicating the test is acceptable to both sensitivity reviewer and test assembler. The sensitivity reviewer also signs the control sheet and charges the time to the project/job for the test.

   b. The test assembler makes the agreed-upon changes, indicating what revisions have been made, and forwards the report form and work folder to the sensitivity reviewer. The sensitivity reviewer signs and dates the report form, indicating the test is acceptable as revised, signs the control sheet, and returns both to the test assembler along with the work folder within 48 hours of receipt. Time is charged to the project/job for the test.

   In cases where there is no resolution, the sensitivity reviewer will record on the report form all of his or her disagreements with the test assembler's responses. The report form and the work folder are submitted to the test sensitivity review coordinator from the test assembler's area. The coordinator submits the material for binding arbitration. In recording his or her position, the sensitivity reviewer should make specific references to relevant sections and pages in the *Guidelines*.

   Both the test assembler and the test sensitivity reviewer write memoranda of explanation to the arbitration panel.

   The test sensitivity review area coordinator requests that the test sensitivity review steering committee chairperson form an arbitration panel.

   All materials go through the test sensitivity review area coordinator to the arbitration panel.

   The arbitration panel gives its decision to the coordinator, who notifies the involved parties.
8. The Test File Library retains the final test sensitivity review report form and the arbiters' decision as permanent components of the work folder.

---

[3] The test sensitivity review report form must not be used for comments or suggestions other than sensitivity concerns. Reviewers are encouraged to make such comments but to write them on a separate sheet of paper.

[4] At any point in the process, the sensitivity reviewer may consult with his or her test sensitivity area coordinator or any other available area coordinator.

Arbitration Process

As soon as it is recognized that arbitration will be required, the test sensitivity review area coordinator should notify the chairperson of the sensitivity review steering committee. It is important that this be done quickly so that an arbitration panel, consisting of three of the five trained arbiters, can be assembled as soon as possible. The chair of the sensitivity review steering committee appoints a chair for the panel and arranges for a meeting room. The panel's decision is due within one week. Time is charged to the project job of the test under consideration.

Since the mandatory sensitivity review occurs during the test editing process, other editorial changes in test material can be made while the sensitivity item is in arbitration. Copy editors should NOT sign off, however, until the control sheet is signed by the test assembler's area director, who notes near the appropriate box: "Decided in Arbitration."

Further procedural steps are as follows:

1. The sensitivity reviewer, test assembler, sensitivity review area coordinator, and test development director sign the test sensitivity review arbitration control sheet. Signatures indicate awareness that the passage/item/test is going to arbitration, not necessarily agreement with either party.
2. The work folder and sensitivity review report form are given to the sensitivity review area coordinator.
3. All arbitration occurs through written material only. There will be no oral arguments by either party before the panel of arbiters.
    a. In a memorandum, the sensitivity reviewer must clearly indicate the nature of the problem(s) and must cite the section(s) and page number(s) of the guideline(s) being violated. A reviewer's inability to find specific references may indicate that the objection is inappropriate.
    b. In a memorandum, the test assembler must clearly indicate the reason(s) for NOT accepting the sensitivity reviewer's recommendations. (Time constraints will not be considered sufficient reason for not changing test material.) The test assembler's statements should explain how and why the test material does NOT violate the guidelines. The test assembler's statements should be documented, including text references to the *Guidelines* and copies of test specifications when appropriate.
    c. Other written material may be solicited by the panel itself.

The arbiters are familiar with the ETS test sensitivity guidelines and have a copy of the procedures available when they meet. The panel can decide one of the following:

1. Passage/item/test is in violation of the guidelines and the material must be changed or dropped.
2. Passage/item/test is not in violation of the guidelines.
3. The guidelines do not address and are not relevant to the problem raised by the reviewer.

In reviewing the passage/item/test, arbiters may discover that another passage/item, not cited by the reviewer, violates a guideline. It is the duty of the arbitration panel to rule on that material as well. The arbitration process is intended to provide a mechanism for resolving disagreements between test assembler and sensitivity reviewer; however, as the fundamental goal of the sensitivity review process is to eliminate offensive material, arbiters would be remiss if they were not to rule on any material brought before them that violates the guidelines.

Once the panel has made its binding decision, the arbitration control sheet, the test sensitivity review report form, and the test assembler's control sheet are signed and returned to the area sensitivity review coordinator. Copies of the arbitration decision and the memoranda written by the test assembler and the sensitivity reviewer are sent by the assembler's area sensitivity review coordinator to the assembler's area director, the steering committee, the test assembler, and the sensitivity reviewer. The sensitivity review area coordinator and the steering committee are also sent copies of the passage or item. The sensitivity review coordinator ensures that any necessary changes are implemented.

PROCEDURES FOR SENSITIVITY REVIEWS OF PUBLICATIONS

Mandatory Review

Sensitivity reviews of all ETS publications other than tests are performed by the editorial staff. Editors, like test reviewers, must go through sensitivity training to be qualified to perform such reviews. If an editor of a publication is also the author of the manuscript, another editor performs the sensitivity review. Editors undertake sensitivity reviews when the manuscript has reached final draft stage, before it is put into production. However, editors are encouraged to review copy informally as early in the editorial process as possible. If copy is changed or added to a manuscript already reviewed for sensitivity and in production, the editor must review the additions for conformity to the ETS sensitivity guidelines. Editors also review publications produced before the most recent guidelines were issued when such publications are scheduled to be reprinted.

Editors are also responsible for reviewing audiovisual publications and artwork proposed for inclusion in publications, using the same procedures described above.

Editorial staff bring sensitivity issues in publications to the attention of the project director. The editor then works with the project director to eliminate questionable or inappropriate copy from the publication. A project director who chooses not to change the copy, due to conflicts with program policies, must reply on the publications sensitivity review form to the editor's objections. If the disagreement continues, the sensitivity reviewer, the project director, and the publications sensitivity review coordinator meet with the division director of the project director's area. The division director serves as a mediator and attempts to resolve the issue(s) to the mutual satisfaction of the sensitivity reviewer and the project director. If the problem is not solved, the publications sensitivity review coordinator notifies the chair of the steering committee, and the dispute goes to arbitration as quickly as possible.

Arbitration Process for Publications

The chair of the sensitivity review steering committee arranges for an arbitration panel, appoints a chair for the panel, and arranges for a meeting room. At this point, the sensitivity reviewer, project director, publications sensitivity review coordinator, and sensitivity review coordinator from the project director's area sign the sensitivity review arbitration control sheet. Signatures indicate awareness that the publication is going to arbitration, not necessarily agreement with either party.[5]

In a memorandum, the sensitivity reviewer must clearly indicate the nature of the problem(s) and must cite the section(s) and page number(s) of the guideline(s) being violated. A reviewer's inability to find specific references may indicate that the objection is inappropriate.

The project director must clearly indicate the reason(s) for NOT accepting the sensitivity reviewer's recommendations. Time constraints will not be considered sufficient reason for not changing material. The project director's statements should explain how and why the material does NOT violate the guidelines. The project director's statements should be documented, including appropriate references to the *Guidelines and Procedures* and copies of relevant specifications when appropriate.

The draft publication, sensitivity review report form, and any explanatory memoranda from the sensitivity reviewer and the project director are given to the publications sensitivity review coordinator, who forwards them to the chair of the arbitration panel. The panel may solicit other written material itself.

An arbitration panel will be convened and a decision rendered usually within one week of notification. Three of the five arbiters will be asked to serve on a panel. Time charges are to be made to the project job of the publication under consideration.

The arbiters have received ETS sensitivity training and have a copy of the procedures available at the meeting. The panel can decide one of the following:

1. Material is in violation of the guidelines and must be changed or dropped.
2. Material is not in violation of the guidelines.
3. The guidelines do not address and are not relevant to the problem raised by the reviewer.

---

[5] All arbitration occurs through written materials only. There are no oral arguments by either party before the panel of arbiters.

In reviewing the material, arbiters may discover that additional areas not cited by the reviewer violate a guideline. It is the duty of the arbitration panel to rule on that material also. The arbitration process is intended to provide a mechanism for the resolution of disagreements between project director and sensitivity reviewer. However, as the major goal of the sensitivity review process is the elimination of offensive material, arbiters would be remiss if they were not to point out and rule on material brought before them that is in violation of any part of the guidelines.

Once the arbitration panel has made its binding decision, its members sign the arbitration control sheet and the publications sensitivity review form and return them to the publications sensitivity review coordinator. Copies of the decision, together with the material under arbitration and the memoranda written by the project director and the sensitivity reviewer, are sent by the publications coordinator to the project director's division head, the steering committee, the project director, and the sensitivity reviewer. The sensitivity review coordinator will ensure that the changes, where necessary, are implemented.

EVALUATION GUIDELINES

The success of the sensitivity review process depends upon the consistent implementation of clear and established policies. It is necessary that reviewers and editors be familiar with all of the guidelines discussed below to ensure that all people and groups are treated fairly in tests and publications and that all test programs and clients are asked to comply with the same standards.

Definitions

Group Reference Questions reflect the multicultural nature of our society and are of two basic types: *representational* and *substantive*.

Representational items

These items test knowledge or skills that are independent of the particular subject matter presented in the stimulus material or in the item itself. Such items are generally found in tests measuring listening skills, reading comprehension, problem-solving in mathematics, writing ability, interpretation of data, and the like. For example, if the purpose of the item is to test whether a candidate knows how to read a bar graph, what the bar graph itself indicates is irrelevant; the same skill can be measured whether the graph compares the number of cars manufactured by different companies, the number of people who are in the various income tax brackets, or the number of Hispanic men and women who have earned doctorates each year during the past decade. Usually, items in this category can be changed without great difficulty to include references to women or minority groups.

Examples

1. Skill — Identification of an error in grammar

Original sentence:

Henry Fielding $\underset{A}{\underline{\text{is widely known}}}$ and highly praised for his novels$\underset{B}{\underline{,}}$ few people $\underset{C}{\underline{\text{realize that}}}$ he $\underset{D}{\underline{\text{established}}}$ the first police force in England. $\underset{E}{\underline{\text{No error}}}$.

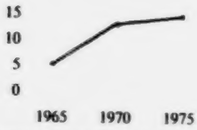
Revised item that includes a representational women's reference:

Gwendolyn Brooks $\underset{A}{\underline{\text{is widely known}}}$ and highly praised for her poetry$\underset{B}{\underline{,}}$ few people $\underset{C}{\underline{\text{realize that}}}$ she has $\underset{D}{\underline{\text{also}}}$ published a novel. $\underset{E}{\underline{\text{No error}}}$.

2. **Skill**—Ability to read a simple chart

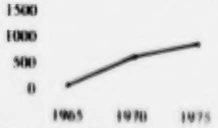
Original chart from which questions were drawn:

Number of Test Development Specialists in State X's Employ



New chart that includes a representational group reference:

Number of Hispanic Americans Holding Professional Jobs in State X's Government



3. **Skill**—Spatial orientation

Original item:

Jim rowed 1 kilometer east and then 1 kilometer south. In what direction would Jim have to row in order to return directly to his starting point?

(A) North (B) Northeast (C) Northwest
(D) Southeast (E) Southwest

Revised item that includes a representational Asian American reference:

Mr. Chyn rowed 1 kilometer east and then 1 kilometer south. In what direction would he have to row in order to return directly to his starting point?

(A) North (B) Northeast (C) Northwest
(D) Southeast (E) Southwest

4. **Skill**—Reasoning

Original item:

If a man who had visited the United States in the 1830s wrote, "People in America were unusually friendly," you would probably give the most credence to his judgment about American people if you also found that

(A) Americans of the time condemned the idea that America was a happy-go-lucky culture
(B) ministers in the 1830s insisted that puritanism was declining
(C) other travelers in the 1830s who came from the same culture as the author had come to the same conclusion
(D) other travelers in the 1830s who came from many different cultures had come to the same conclusion as the author
(E) the first American social club was founded in the 1830s

Revised item that includes a representational women's reference:

If a woman who had visited the United States in the 1830s wrote, "Unmarried women in America were unusually emancipated," you would probably give the most credence to her judgement about these women if you also found that

(A) social psychologists in the 1980s contend that women in the United States are more emancipated than women in most societies

(B) United States writers of novels in the 1830s described some women characters who refused to follow established rules of conduct

(C) in the 1830s, another traveler, who came from the same culture as the author, had come to the same conclusion

(D) in the 1830s, men and women travelers, who came from many different cultures, had come to the same conclusion as the author

(E) the first suffragist newspaper in the United States was founded in the 1830s

Definition

Substantive item

Substantive items test particular kinds of knowledge. These items are usually found in tests meant to measure knowledge gained in a particular course of study in a particular discipline. Substantive items related to the concerns of minority groups and women are included in the test according to the requirements of the test specifications, which, of course, are intended to reflect what is being taught in the discipline. Some of these items may cover subjects that can be expected to arouse negative emotional reactions in certain subgroups of the population and thus would not be appropriate subjects to cover in representational items. For example, a test in American history would probably deal with slavery; a test for nurses might include items about sickle-cell anemia or Tay-Sachs disease. All such items should be reviewed for sensitivity concerns in light of the purpose of the test, the population taking it, and the curriculum it is designed to test.

Evaluation Requirements

All questions, including group reference questions, and where applicable entire test sections or tests, are evaluated from a number of perspectives.

Cognitive/Affective

These two dimensions apply to all group reference questions. The cognitive dimension deals with the factual basis of questions, i.e., whether the information in the question is accurate. The affective dimension reflects the positive or negative feelings the question may evoke from various segments of the testing population. There are four possible combinations of these two factors, illustrated by the following chart.

COGNITIVE

AFFECTIVE

	Factual +	Erroneous −
+ positive	a	b
− negative	d	c

Category "a"
ACCEPTABLE: Category "a" represents the ideal situation—the group reference is both factual and affectively positive.

Example

The economic health of the Osage took a dramatic turn for the better when
(A) they succeeded in producing an especially fine variety of cotton
(B) pooled tribal resources provided the capital to establish a pencil factory
(C) oil was discovered on their reservation
(D) high-fashion designers displayed an interest in their finely crafted jewelry
(E) concern for the environment led to a general interest in handcrafted goods

Category "b"
UNACCEPTABLE: Category "b" questions, while evoking positive feelings on the part of referenced groups, are not factual. Such questions frequently result from the intentional efforts of the person writing the question to correct a perceived injustice to a minority group and often represent a narrow ideological perspective. Additionally, these questions tend not to have clear-cut correct answers. *In most cases, unacceptable questions can be salvaged by revision.*

Example

Which of the following groups has been most successful in obtaining progress for the Black community?
(A) The Urban League
(B) The Black Panther Party
(C) The Deacons for Defense
(D) The National Association for the Advancement of Colored People

Discussion

The problem here is with the question itself. Unfortunately, it is unanswerable as written. What is meant by "most successful"? At what? What type of progress? The writer clearly had good intentions. The objective was to include positive material on the minority experience. A question of this type can be rewritten in several ways. For example: Which of the following groups emphasizes progress through alliances with the business community? The answer is the Urban League. Or: Which of the following groups is the oldest? The answer is the NAACP. Both questions as rewritten have a factual as opposed to a subjective answer. Notice that as used in option (D) the word Colored is acceptable and appropriate here.

Category "c"

UNACCEPTABLE: This third set represents the worst case. These questions are not factual, and they generate negative feelings on the part of referenced groups.

Example

All of the following groups have retained some of their original cultural roots EXCEPT the

(A) Swedish Americans
(B) Italian Americans
(C) Black Americans
(D) American Indians

Discussion

The author of this question intended for the answer to be (C). However, one school of thought on this issue traces the roots of Black-American culture clearly back to Africa. Black Americans who support this alternative viewpoint would react negatively to the question as written. Therefore, the question should be dropped or reworded to read: According to E. Franklin Frazier (or some other proponent), which of the following groups has not maintained vestiges of its original cultural heritage?

Category "d"

UNACCEPTABLE: Questions that fall into this fourth group often lead to a controversy that is difficult to resolve. Although such questions are based on fact, they generate negative feelings on the part of referenced groups. For instance, a question that emphasizes high birth rates in certain nations has a factual basis, but it may evoke negative feelings in Americans who can trace their roots to these nations, and it reinforces negative stereotypes.[*]

Example

All of the following factors account for the use of English as the official language of the United States EXCEPT:

(A) It is required by a constitutional amendment.
(B) It is the primary language of instruction in public schools.
(C) It is the key to the "Americanization" of non-English-speaking immigrant groups.
(D) It is usually necessary for career success.
(E) It prevents the emergence of balkanization and separation.

Discussion

Even though choices B-E are true, the question can offend American citizens who are not native speakers of English, as well as recent immigrant groups. Choices C and E show an intolerance of other cultures.

---

[*] In exceptional instances, material of this nature may be unavoidable. See section on context considerations.

Example

Percent of Female-Headed Families in the United States in 1960 by Annual Income, Race, and Place of Residence

	Rural	Urban	Total
	Percent	Percent	Percent
Black Population			
Under $3,000	18	47	36
$3,000 and over	5	8	7
Total	14	23	21
White Population			
Under $3,000	12	38	22
$3,000 and over	2	4	3
Total	4	7	6

The data in the table above indicate that in the United States in 1960 female-headed families were more common

(A) in rural areas than in urban areas
(B) among Whites than among Blacks at the same income level
(C) among poor Whites than among nonpoor Blacks
(D) among the poor than the nonpoor only in urban areas
(E) among Blacks than Whites in urban areas but not in rural areas

Discussion

This item was written for a *general background test*. It is unacceptable in that it

- presents a negative picture of the minority group discussed in the item,
- may arouse negative feelings in test takers,
- is not measuring knowledge of information essential in a discipline,
- is intended for a general population (not students of a particular curriculum), and
- does not treat the subject with as much sensitivity as it could be treated.

Example

Population growth rates tend to be highest

(A) among the poor
(B) in industrial countries
(C) in areas with rich food supplies
(D) when birth rates are low and death rates are high
(E) when a nation undergoes a period of severe economic depression

Discussion

This item was written for a test intended for postgraduates with a special interest in political affairs, economics, and social structures in the United States and throughout the world. In fact, candidates taking the test are expected to demonstrate more than average competence in answering questions in these areas. Given the special purpose of the test, the population, and the treatment of the subject in the question, the item is acceptable for the test.

Controversial Material

Highly controversial issues, such as legalized abortion or hypotheses about genetic inferiority, must not be included in any test question unless such issues are both relevant and essential to the content validity of the test. If such material is to be used, the question must be constructed to indicate clearly its relationship to the content validity of the test. Several methods for accomplishing this within the sensitivity guidelines are:

- Identify the source. For example, one could begin a question with "According to (source) . . ." or "In the opinion of (source). . . ."
- Phrase the question in such a way as to require an in-depth knowledge of the subject matter.
- Balance the first controversial question with another that either refutes the first or presents an alternative point of view.

Examinee Perspective

All group reference questions are reviewed from the perspective of test takers who may not have access to the correct answers. When an examinee must know the correct answer to prevent a question from reinforcing negative attitudes or stereotypes, the question should be revised or rejected. This is because examinees who select a wrong option are not routinely informed that their response was incorrect. Thus their belief in the legitimacy of a negative attitude may be reinforced.

In evaluating perspective, the sensitivity reviewer must recognize that there will be instances, particularly in content-based tests, where negative statements must appear. For example, negative statements are to be expected in literature tests, especially material dealing with satire or irony, where the author's statements may address either individuals or groups. Similarly, in sociology, history, or economics tests, conflicts or development patterns frequently require knowledge about, or interpretations of, social and/or cultural documents and concepts that may seem offensive to individuals or groups. In such instances, the test assembler must be able to demonstrate that a potentially offensive option is a legitimate part of 1) accurately interpreting a required kind of stimulus material or 2) accurately demonstrating an understanding of the knowledge base of a particular discipline. The assembler's inability to demonstrate such points will suggest that the distracter should be revised.

Balance

In general, the sensitivity reviewer should determine whether there is a suitable balance of multicultural material in *final forms* of a test or test section. In tests that largely test skills, such as mathematics aptitude tests and writing ability tests, the numbers of references to males and females in items that refer to people should be approximately equal. Such tests should also contain references to one or more minority groups, at least meeting the test's own specifications on multicultural representation. If such a test consists entirely of a small number of passages, such as some reading comprehension tests, balance requirements should be applied less stringently—for example, if one out of three passages focuses on either women or a minority group, the test's balance is acceptable.

Tests that largely assess content should meet their own specifications on sex-referenced and multicultural material. If the test's specifications do not refer to women and minority groups, ETS's corporate guideline requiring "the inclusion of material reflecting the cultural background and contributions of major population subgroups" should be followed. For example, women and minority groups could be mentioned in items that test skills (for example, as the topic of a graph in a graph-reading item in an economics test).

Tests that assess a mixture of content and skills should be evaluated individually, applying the spirit of this guideline. Such tests may include curriculum-based skills tests, such as the interpretation and analysis of literature, and occupation-based skills tests, such as police officers' examinations.

In all tests, it is desirable to refer to more than one minority group, rather than focusing all items on a single minority group.

Because many programs use pretesting to build and augment question pools for the assembly of scored tests meeting strict content and statistical specifications, the sensitivity reviewer cannot require that a *pretest* be balanced in its representation of either women or minorities if the pretest specifications do not specifically require such material. Notation of pretest specifications should be made by the test assembler on the test sensitivity review report form.

In judging the balance of a test, the sensitivity reviewer should consider not only the numerical balance of sex and minority representation but also a more holistic appraisal of the overall impression that is made by the test's references to women and minority groups. The ways in which men, women, and minority group members are portrayed and the strength of the various references are among the factors to consider in making such a holistic appraisal.

For computerized, self-selecting, or branching tests, the entire pool of items should be reviewed before the system is used. At that time, the pool should be evaluated to see if it contains an acceptable balance.

References to persons with disabilities are not part of the balance requirement for any test.

Stereotyping

Sensitivity reviews must ensure that no test implies that a population group is culturally or biologically inferior or superior to any other group. Thus, the review should ensure that no material contains language or symbols that reinforce offensive stereotypes. Such stereotypes generally suggest the physical (e.g., height, weight, attractiveness, strength) or psychological (e.g., intelligence, ethics, emotions, behavioral patterns) inferiority of a particular group in some characteristic considered desirable by the majority culture. For example, material that refers to the alleged predominance of Italian Americans in organized crime implies that Italian Americans are dishonest. Occasionally, an offensive stereotype implies a superiority of one group over another. For example, many would view as offensive a question that implies that males are better drivers than females. Material judged to contain language or symbols that reinforce offensive stereotypes is not acceptable. (See Appendix A for examples of offensive stereotypes.)

It is also important to avoid stereotyping women or a minority group by portraying them in only one role, especially if it is a stereotypical role. Instead, they should be portrayed engaging in many different activities. For example, if a woman is engaged in a traditional activity like child-rearing in one item, it is desirable to have one or more items in the test in which women are engaged in less traditional activities, such as working as a lawyer or business executive.

In evaluating stereotypes, the sensitivity reviewer must recognize that there will be instances where stereotypes are likely to appear as part of the content-related material of a test. For example, there may be instances where a social worker must know common stereotypes in order to deal with social problems or where a historian must be aware of stereotypes in order to accurately interpret historical documents. Here, as in evaluating perspective, the assembler must be able to demonstrate that the presence of a stereotype and the test taker's ability to recognize and interpret it are required by the discipline.

Caution Words and Phrases

Sensitivity reviews should reflect the fact that even words with legitimate uses can sometimes appear in contexts that make them unacceptable. Through experience, sensitivity reviewers have learned that certain key words and phrases often accompany sensitive material. Thus, although the use of these words and phrases is proper and legitimate, the appearance of these words indicates that the material requires special attention because of an increased potential for offensiveness. Examples of such words are *lower-class*, *discrimination*, and *race*. (See Appendix B for a more comprehensive list of examples.)

Special Review Criteria for Women's Concerns

Sensitivity reviewers should seek to identify and eliminate all language that discriminates on the basis of sex. (See Appendix C for detailed requirements.)

Special Review Criteria for References to People with Disabilities

Sensitivity reviewers should seek to identify and eliminate all language that discriminates on the basis of physical or mental disabilities. (See Appendix D for detailed requirements.)

Underlying Assumptions

Sensitivity reviewers should attempt to eliminate ethnocentric or gender-based beliefs and prejudices. An underlying assumption is a subtle secondary premise that reflects an individual's ethnocentric beliefs. Unacceptable underlying assumptions include the following:

- That a group is deserving of a particular fate.
- That a group is by nature dependent on help from another group.
- That a group lacks or has an excess of any given quality fairly common to humans.
- That what may be a norm only in Western culture is "truth" or that European civilization is "better" than (as opposed to "different" from) other civilizations.
- That a causal link exists between any group and poverty, crime, intelligence, and the like.

Context Considerations

Sometimes the use of sensitive material is unavoidable. There are four areas in which this occurs with some frequency.

1. *Historical Domain*: In order to measure an individual's knowledge of history, it may sometimes be desirable to quote from material written during a period when social values differed markedly from today's. For example, a passage describing the conditions of Southern Black people during the reconstruction period may include the term "colored people" or "Negro." While it is desirable to avoid the use of such material where possible, the sensitivity issues must be judged in the overall context in which they are presented.
2. *Literary Domain*: Material that is designed to measure an individual's knowledge of literature or that quotes from works of literature often contains similar problems. For example, many passages from material written before the 1970s may include constant use of the so-called "generic he," a style that was considered editorially correct until recently. Similarly, passages that deal with cultures other than the majority culture may vary in purpose or in methods of discussing ethnic ideals or attitudes. In all such instances, the stimulus material and items must be reviewed carefully.
3. *Legal Domain*: Material drawn from legal sources may sometimes deal with sensitive areas. For example, real estate tests may contain references to federal, state, or local laws governing discrimination in the mortgage rights of EEO classes.
4. *Health and Social Sciences Domains*: Certain examinations in these domains (including health professions, social work, and civil service) require knowledge of information that may be considered sensitive in other contexts. For example, in nursing tests it may be necessary to test one's knowledge of the predominance of sickle-cell anemia among Black people or Tay-Sachs disease among Jewish families. Social work and the civil service require knowledge on how to approach problems and/or counsel people in a wide variety of social and cultural contexts.

Inclusion of potentially sensitive material depends on the content of the entire test or publication. Given an appropriate context, use of certain material may be justifiable. It is important to recognize that many subject-matter tests must include information and concepts that have a great potential for raising sensitivity issues. The test assembler and the sensitivity reviewer are responsible for working together to develop test material that covers necessary subject matter (such as slavery, the Japanese-American internment during World War II, ethnic components of social problems) in a theoretically balanced, sensitive, and objective manner.

Elitism, Ethnocentricity, and Related Problems

To eliminate concepts, words, phrases, or examples that may upset or otherwise disadvantage a test taker, every effort is made not to include expressions that might be more familiar to members of a particular social class or ethnic group than the general population, such as "soul food" and "trust fund," unless they are defined or knowledge of them is relevant to the purpose of the test. Words and sentence constructions that could have different meanings to different ethnic or geographic groups must be avoided. Care must also be taken to assess the appropriateness of dialect, slang, and non-English words and phrases, such as "bairn," "stickball," and "maven," which tend to be more familiar to certain ethnic, geographic, or other subgroups of English speakers.

APPENDIX A

GUIDELINES FOR RECOGNITION OF UNACCEPTABLE STEREOTYPES

This appendix lists some of the stereotypes that have been identified by members of major population groups. When reviewing these examples, it is important to understand that they are not intended to serve as an exhaustive list of all possible stereotypes but rather to illustrate some of the more commonly encountered ones.

Although some words in this appendix may seem dated or may appear infrequently in contemporary texts, they are found in many sources (such as literary passages, historical documents and cartoons, and popular publications) that provide stimulus material for test questions. Such words are listed here in order to remind the reviewer that they, and words like them, must always be carefully evaluated, regardless of the context in which they appear.

1. No population group should be depicted through language or symbols as superior or inferior with regard to:
    - contribution to society
    - education
    - emotional stability
    - honesty
    - industriousness
    - intelligence
    - leadership ability
    - morality
    - physical appearance
    - physical capabilities
2. No population group should be depicted through language or symbols as fixated on instant gratification (unable to plan for the future).
3. No population group should be depicted as unable to mix socially with other groups.
4. No population group should be depicted as superior or inferior in its social institutions, social organizations, or social structures.

Examples of Unacceptable Stereotypes

1. That Asian/Pacific Americans:
    - are only suited for certain vocations and professions (e.g., food service work, laundry work, mathematics);
    - speak "pidgin" English;
    - are short, skinny, and wear glasses;
    - subsist on chop suey, fried rice, herbal tea;
    - live or prefer to live in ethnic neighborhoods (e.g., Chinatown, Little Seoul);
    - are predominantly refugees;
    - marry in accordance with family wishes or as a result of a prearranged agreement between families.
    - practice polygamy;
    - favor sons over daughters and first sons over all other siblings;
    - have little regard for human life;
    - use narcotics, particularly opium and its derivatives;
    - require women to be passive and submissive;
    - all share the same basic culture (as opposed to recognizing substantial cultural variations that exist in the heritages of Asian/Pacific Americans).

2. That Black Americans:
   - are only suited for certain vocations and professions (e.g., sports, music, teaching);
   - are less prepared or less adequate as professionals;
   - comprise the majority of individuals receiving welfare support;
   - (males) often desert their families;
   - are not punctual;
   - frequently engage in civil disorders and looting;
   - live exclusively in depressed urban areas;
   - are licentious (overpopulate, routinely engage in sexual relations at a young age, etc.);
   - are unaware of their African heritage;
   - gamble excessively;
   - drink excessively;
   - have an inherently superior sense of rhythm;
   - speak "Black" language;
   - excel in physical as opposed to intellectual endeavors.
3. That Hispanic Americans:
   - are only suited for certain vocations and professions (e.g., service work, agricultural work);
   - are licentious (overpopulate, routinely engage in sexual relations at a young age, etc.);
   - are violent or bloodthirsty (bullfighting, revolutionary, etc.);
   - receive a disproportionately high percentage of welfare support;
   - speak dialects of Spanish unintelligible to other Hispanic groups;
   - are not punctual and frequently procrastinate (mañana attitudes);
   - (men) physically dominate women (macho attitudes);
   - are all alike as opposed to recognizing cultural differences (e.g., Puerto Rican, Cuban, Chicano, etc.).
4. That Native Americans/American Indians:
   - are unable to handle alcoholic beverages;
   - are "closer to nature" than other Americans;
   - live in teepees and/or slums;
   - lack the ability to deal with modern technology;
   - lack the ability to deal with intellectual and academic endeavors;
   - are unusually hostile, violent, or apathetic;
   - are all alike (as opposed to recognizing substantial variations among and within the Indian Nations);
5. That women:
   - are only suited for certain vocations and professions (e.g., elementary school teaching, nursing, secretarial, librarian);
   - are less prepared or less adequate as professionals than men;
   - are weak, fragile, or passive;
   - are overly emotional or hysterical (panic in crises);
   - are disorganized, illogical, or scatterbrained;
   - frequently engage in gossip;
   - compete with each other;
   - lack basic mechanical ability (e.g., can't drive a car or fix a leaky faucet);
   - lack ability to excel at any activity (music, science, etc.);
   - are overly concerned with their physical appearance;
   - are pushy;
   - lack qualities of leadership (i.e., self-confidence, ambition, and assertiveness);
   - lack basic ability in mathematics.
6. That persons with disabilities:
   - are helpless or less able than others who take care of themselves;
   - are to be pitied or patronized;
   - are nonproductive members of society;
   - are in need of government assistance.

Appendix B

CAUTION WORDS AND PHRASES

Experience has shown that the following words and phrases frequently accompany sensitive material. While the vast majority of these words and phrases are themselves legitimate and are often used appropriately, they tend to indicate an increased potential for the presence of offensive material.

Although some words in this appendix may seem dated or may appear infrequently in contemporary texts, they are found in many sources (such as literary passages, historical documents and cartoons, and popular publications) that provide stimulus material for test questions. Such words are listed here to remind the reviewer that they, and words like them, must always be carefully evaluated, regardless of the context in which they appear.

1. Caution words and phrases with regard to all population groups:
   - affirmative action
   - backlash
   - backward, backwardness
   - barbarian, barbaric
   - birthrate
   - civil disorder
   - civilized
   - class, lower class, middle class, upper class
   - colonialism, colonized
   - crime, criminal, crime rates
   - culture, cultural bias,
   - *culturally deprived
   - *culturally disadvantaged
   - deficient
   - deprived
   - developing nation
   - deviance, deviant behavior
   - dialect
   - disadvantaged
   - discrimination
   - emotional, emotionalism
   - environment
   - equality
   - freedom
   - gangs
   - genetic, genetic inferiority, genetic superiority
   - ghetto
   - ignorant
   - illegitimate
   - illiteracy, illiterate, illiterates
   - inequality
   - inferior
   - inner city
   - instant gratification
   - intelligence, intelligence test
   - juvenile delinquency
   - masses, the masses
   - melting pot
   - minority
   - nonwhite
   - single-parent family
   - physical type, physical capabilities, physical characteristics
   - preferential treatment
   - primitive
   - promiscuous
   - race, racism
   - riot
   - ritual
   - social class, social development
   - socioeconomic
   - Third World
   - uncivilized
   - underprivileged
   - *underdeveloped nations
   - uneducated
   - urban
   - violent, violence
   - welfare
2. Caution words and phrases with regard to Asian/Pacific Island Americans:
   - Asian American(s)
   - *Chink (demeaning abbreviation of Chinese)
   - Chinaman, Chinamen
   - Chinawoman, Chinawomen
   - *Far East
   - *Jap, Japs
   - Japanese
   - *Orient
   - **Oriental(s)
   - Pacific Islander(s)

Note: The distinct terms *Asian American*, *Pacific American*, and *Asian/Pacific Island American* should be used according to accuracy and appropriateness.

---

* These are generally unacceptable terms.

**Whenever possible avoid using these terms as nouns. It is preferable to use them as adjectives, i.e., Asian Americans or Black people.

3. Caution words and phrases with regard to Black Americans:

Africa, African
African(s), Afro-American(s)
Black
**Blacks
Black Americans
busing
* colored, colored people
desertion, desertion rates
integrate, integration
jungle
native
* Negro, Negroes
people of color
primitive
segregate, segregation
slaves, slavery
South Africa, South African(s)
tribe, tribal

4. Caution words and phrases with regard to Hispanic Americans:

barrio
bilingual
Chicano(s)
Cuba, Cuban
Cuban(s), Cuban American(s)
extended family
Hispanic
**Hispanics
Hispanic American(s)
**Latin(s), Latin American(s)
Latino
macho, machismo
* Mex
Mexico, Mexican
Mexican American(s)
nation, nations
New Rican
Puerto Rico, Puerto Rican
Puerto Rican(s), Puerto Rican American(s)
Spanish, Spanish American(s)
* Tex-Mex

5. Caution words and phrases with regard to Native Americans/American Indians:

Aleut(s) (Use this form instead of Eskimo.)
American Indian(s)
* Eskimo(s)
Indian
Inuit(s), Innuit(s) (Use this form instead of Eskimo.)
native
Native American(s)
* redman, redmen
reservations
treaty, treaty privilege
tribe, tribal

Note: The terms *Native American* and *American Indian* are both acceptable and may be used independently, as appropriate.

---

* These are generally unacceptable terms.
**Whenever possible avoid using these terms as nouns. It is preferable to use them as adjectives; i.e., Asian Americans or Black people.

6. Caution words and phrases with regard to women and men (see also Appendix C):

- * better half
- boy(s), boyish
- * coed (as a noun)
- * distaff side
- domineering
- females, feminine, feminist
- frivolous
- gender
- he, his, him
- * housewife
- homemaker
- hysterical
- lady, ladyish
- libber, women's libber
- maid, maiden
- male(s), masculine
- man, manly, manhood, men
- matriarch
- Miss, Mrs., Ms.
- mother, mother-in-law, grandmother
- nosey
- old maidish
- patriarch
- picky
- pushy
- woman, womanly, womanhood
- women
- sex, sexes, sexy
- she, her, hers
- stubborn

7. Caution words and phrases with regard to persons with disabilities (see also Appendix D):

- * afflicted/inflicted
- * crippled
- * deaf and dumb
- * deformed
- * drain/burden
- normal
- patient
- retarded
- **the deaf, the blind, the handicapped
- * wheelchair bound, confined/restricted to wheelchairs

---

* These are generally unacceptable terms.

**Avoid using these terms as nouns. It is preferable to use them as adjectives; e.g., Asian Americans or Black people or deaf students.

Appendix C

SPECIAL REVIEW CRITERIA FOR WOMEN'S CONCERNS[1]

1. Women must not be described by physical attributes when men are being described by mental attributes or professional position. Irrelevant references to a man's or a woman's appearance, charm or intuition are not acceptable.
2. In descriptions of women, a "patronizing" or "girl-watching" tone is not acceptable, nor are sexual innuendoes, jokes, or puns. Examples of unacceptable practices are: focusing on physical appearance (a buxom blonde), using special female-gender word forms (poetess, aviatrix, usherette), and treating women's issues as humorous or unimportant. The following list identifies a number of generally unacceptable words and phrases and presents one or more acceptable substitutes for each case:

Unacceptable	*Acceptable*
the fair sex; the weaker sex; the distaff side	women
girl, as in: I'll have my girl check that.	I'll have my secretary (or my assistant) check that. (Or use the person's name.)
lady used as a modifier, as in *lady lawyer*	*lawyer* (A woman may be identified simply through the use of pronouns, as in: The lawyer made her summation to the jury. When gender modifiers are required, use *woman* or *female*, as in: a course on women writers, or the airline's first female pilot.
the little woman; the better half; the ball and chain; and other such colloquialisms	wife, spouse
female-gender word forms, such as *authoress* or *poetess*	author, poet (Some words like heroine or actress can be used if they seem appropriate in the given context.)
female-gender or diminutive word forms, such as *suffragette*, *usherette*, *aviatrix*	suffragist, usher, aviator (or pilot)
libber (a put-down)	feminist
sweet young thing	young woman; girl
coed (as a noun)	student
housewife	*homemaker* for a person who works at home, or rephrase with a more precise or more inclusive term
career girl or career woman	Identify the woman's profession: attorney Ellen Smith; Maria Sanchez, a journalist or editor or business executive or doctor or lawyer or agent.
cleaning woman, cleaning lady, or maid	housekeeper; house or office cleaner

3. In descriptions of men, especially men in the home, references to general ineptness are not acceptable. Men should not be characterized as dependent on women for meals, clumsy in household maintenance, or foolish in self-care.
4. Women must be treated as part of the rule, not as the exception. Generic terms, such as doctor and nurse, are assumed to include both men and women, and modified titles such as *woman doctor* or *male nurse* are not acceptable. Stereotyping work activities as "woman's work" or a "man-sized" job is not acceptable.

---

[1] Adapted from McGraw-Hill's *Guidelines for Equal Treatment of the Sexes*. Used with the permission of McGraw-Hill Book Company. (See also Appendix B, Section 6.)

5. Women should be spoken of as participants in any action, not as passive bystanders. Terms such as *pioneer*, *farmer*, and *settler* must not be used as though they apply only to adult males. Examples follow:

Unacceptable	*Acceptable*
Pioneers moved West, taking their wives and children with them.	Pioneer families moved West. or Pioneer men and women (or pioneer couples) moved West, taking their children with them.

6. Women must not be portrayed as needing male permission in order to act or to exercise their rights. Example:

Unacceptable	*Acceptable*
Jim Weiss allows his wife to work part-time.	Judy Weiss works part-time.

7. The word *man* has long served to denote both a person of male gender and humanity at large. To many people today, however, the word *man* is so closely associated with the first meaning (a male human being) that it is no longer considered broad enough to be applied to a person of either gender. Therefore, alternative expressions must be used in place of *man* (or derivative constructions used generically to signify humanity at large). The following list identifies acceptable alternatives for *man*-words:

Man-word	*Preferred Alternative*
mankind	humanity, human beings, human race, people, humankind
man's achievements	human achievements
If a man drove 50 miles at 60 mph...	If a person (or driver) drove 50 miles at 60 mph...
the best man for the job	the best person (or candidate) for the job
manmade	artificial, synthetic, constructed, fabricated
manpower	human power, human energy, workers, work force, human resources, personnel
grow to manhood	grow to adulthood, grow to manhood or womanhood

8. Use of the so-called "generic he" is unacceptable. Here, as elsewhere, historical context and/or direct quotations must be considered when evaluating material.

Passages chosen for reading comprehension items in admissions tests may not use the "generic he." Tests or test sections composed of discrete items have the following possibilities:

A) If the item has several references, balance the use of "he" and "she" within the item.

B) Change "he" to a specific name: Sam, Jim, etc. Change other items to Jane, Cheryl, etc. Then balance the items throughout the test or test section.

Finally, note that a stem like "A man drove 50 miles . . . ." is not a "generic he" item. Items like this need only be balanced with items like "A woman invested . . . ." Examples of other alternatives follow:

Unacceptable	*Acceptable*
The average American drinks his coffee black.	The average American drinks black coffee.

Replace the masculine pronoun with *one*, *you*, *he* or *she*, *her* or *they*, or *people* as appropriate. Alternate male and female expressions and examples to establish a balance within an item.

Example:

Unacceptable	*Acceptable*
I've often heard supervisors say, "He's not the right man for the job," or, "He lacks the qualifications for success."	I've often heard supervisors say, "She's not the right person for the job," or, "He lacks the qualifications for success."

9 Occupational or activity terms ending in *man* are not acceptable when they can include members of either sex. Exceptions can be made for references to a particular person. Examples:

Unacceptable	*Acceptable*
congressman	member of Congress; representative (but Congressman Koch and Congresswoman Holtzman are acceptable)
chairman	chair, chairperson, the person presiding at (or chairing) a meeting, the presiding officer, head, leader, coordinator, moderator (Also acceptable are Chairwoman Shirley Chisholm and Chairman Mao.)

(Note that "Chairman John Doe and Chairperson Jane Doe" is not an acceptable combination, since *man* and *person* are not parallel.)

businessman	business executive
fireman	firefighter
mailman	mail carrier, letter carrier
salesman	sales representative, salesperson, sales clerk
insurance man	insurance agent
cameraman	camera operator
foreman	supervisor

10 Test items that assume all test takers to be male are unacceptable.

Unacceptable	*Acceptable*
you and your wife	you and your spouse
when you shave in the morning	when you brush your teeth (or wash up) in the morning

11 Parallel language must be used for women and men.

Unacceptable	*Acceptable*
the men and ladies	the men and the women, the ladies and the gentlemen, the girls and the boys
man and wife	husband and wife

(Note that *lady* and *gentleman*, *wife* and *husband*, and *mother* and *father* are role words. *Ladies* should be used for women only when men are being referred to as *gentlemen*. Similarly, women should be called *wives* and *mothers* only when men are referred to as *husbands* and *fathers*. Like a male shopper, a woman in a grocery store should be referred to as a *customer* and not as a *housewife*.)

12 A woman must be referred to in a manner that is parallel with references to a man. Both should be called by their full names, by first or last name only, or by title. Examples:

Unacceptable	*Acceptable*
Bobby Riggs and Billie Jean	Bobby Riggs and Billie Jean King
Billie Jean and Riggs	Billie Jean and Bobby
Mrs. King and Riggs	King and Riggs, Ms. King (because she prefers Ms.) and Mr. Riggs
Mrs. Meir and Moshe Dayan	Golda Meir and Moshe Dayan

13 Women should be identified by their own names (e.g., Indira Gandhi). They should not be referred to in terms of their roles as wife, mother, sister, or daughter, nor should they be identified in terms of their marital relationships unless paired up with similar references to men or such references are basic and necessary to effective measurement.

14. Pronouns must not be linked with certain work or occupations on the assumption that the worker is always (or usually) female or male. Examples

Unacceptable	*Acceptable*
the consumer or shopper...she	consumers or shoppers...they
the secretary...she	secretaries...they
the breadwinner...his earnings	the breadwinner...his or her earnings or bread winners...their earnings

15. Males should not always be first in order of mention

Appendix D

SPECIAL REVIEW CRITERIA FOR REFERENCES TO PEOPLE WITH DISABILITIES[*]

Sensitivity reviewers should be particularly aware of the ways in which people with disabilities are portrayed. People and their worth as individuals should be emphasized, not the disabling conditions they may have. Referring to people *as* their conditions is demeaning and inaccurate.

All terms that have negative connotations or that reinforce negative judgments (e.g., *crippled man* or *crazy woman*) should be replaced with terms that are as objective as possible. No one who has a disability should be pictured as helpless or pitiful. People who have disabilities may be parents, teachers, business owners, leaders in their communities—in short, responsible, productive members of society who are neither to be pitied nor patronized.

For general publications, as well as for tests in which sentences and reading passages contain general information but are not testing knowledge of that information (e.g., SAT and GRE sentence-completion items), it is important to watch for labels attached to people. Identifying a computer programmer as paraplegic or an artist as learning disabled, for instance, is probably gratuitous and irrelevant to the programmer's or the artist's ability to function. On the other hand, it may be acceptable in a test to have a reading passage that describes how one person sucessfully manages a particular disability.

Although there is considerable agreement among organizations that represent or are concerned about particular groups regarding language usage and appropriate terminology, in both instances differences of opinion still exist. Sometimes usage that ETS would prefer to avoid may be part of the historic title of an organization, e.g., the American Council of the Blind. In this instance, the word *blind* is used as a noun instead of an adjective, which is the generally preferred use. Sometimes it is the term itself that is no longer appropriate (e.g., *mental deficiency*, *afflicted*). If an association, journal, or publication has such a term in its name (e.g., the American Association on Mental Deficiency) then one must use the correct name of the organization. However, the use of these terms should be avoided where it is appropriate to do so.

The following unacceptable terms and the preferred alternatives are meant to be guidelines—*not* absolute, inflexible standards. Tests or other publications that deal specifically with teaching, diagnosis, or treatment may require using terms on the unacceptable list in order to convey technical information. If so, the test assembler should check the "special considerations" box on the front of the test sensitivity review report form and note that the test contains specialized material and explain for whom the test is intended. A publications editor should note the specialized material on the publications sensitivity review form.

Generally Unacceptable	*Preferred Alternatives*
the use of a handicapping condition as a noun, e.g., *the* deaf, *the* blind, *the* handicapped	use as adjectives: a deaf student; a blind child; handicapped people
afflicted/afflicted with/ afflicted by/affliction	person who has ----; people who are affected by ----
confined to/restricted to a wheelchair/wheelchair bound	person who uses a wheelchair; person who gets around by wheel-chair; wheelchair user

---

[*] In part derived from literature issued by the Gilbert M. and Martha H. Hitchcock Center for Graduate Study and Professional Development, the University of Nebraska-Lincoln, School of Journalism, the Ontario March of Dimes brochure, the National Easter Seals Guidelines, the Cerebral Palsy Foundation, and a guidebook published by the International Association of Business Communications.

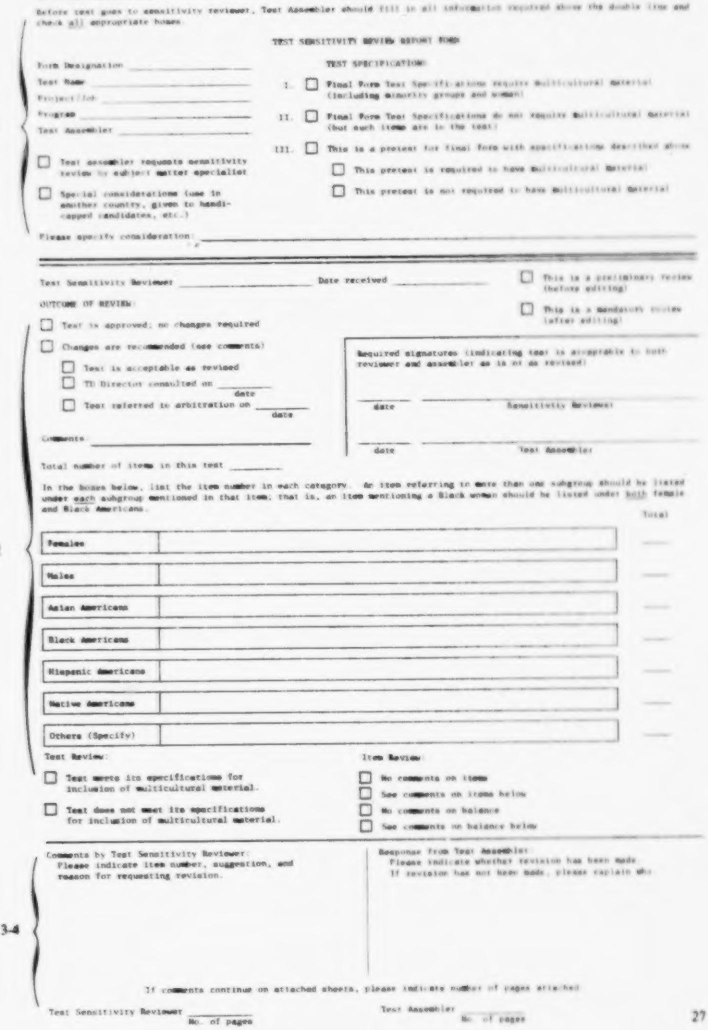
cripple/crippled	person who has a physical disability; physically disabled people
deaf and dumb	people who cannot hear or speak
diagnose/diagnosed	correct only to describe a condition, *not* a person; the condition was diagnosed as ----
disease	use the word condition or specify the name of a condition, such as a person who has multiple sclerosis
drain/burden	person who has a condition that requires increased (or additional) responsibility (or care or intensive care)
inflicted with/inflicted	caused by ----; disabled by ----;
normal	people without disabilities, nondisabled people; nonhandicapped person
patient (noun)	use only to refer to a person who is being treated by a physician at home or who is in a hospital
retarded	See note 3 below
victim/victim of	person who has ----; people who experience ----
blind as a bat/crazy/crip/deformed/dumb/freak/gimp/insane/pitiful/poor/unfortunate	these terms and others like them should NEVER be used

Additional Notes

1. There are guide dogs and seeing-eye dogs for persons who are blind and hearing-ear dogs for persons who are deaf.
2. When people who are deaf communicate by the use of their hands, they may be described as signing. People are described as *interpreting* when they render what someone is saying into sign language for a deaf person.
3. In addition to the general guidelines discussed, the NTE Education of Mentally Retarded Students test committee has made several more decisions about appropriate and current terminology in the field of mental retardation. Among them:

- Use *mentally retarded* rather than *retarded* alone--e.g., mentally retarded students rather than retarded students.
- To specify degrees of mental retardation, use the following:
  mildly retarded (educable) student
  moderately retarded (trainable) student
  severely and profoundly retarded student
- Use the term *mildly mentally retarded* in place of *cultural-familial retarded*.
- Use the term *Down syndrome* rather than *Down's syndrome* (in keeping with new terminology in the 1983 AAMD *Classification in Mental Retardation*).
- Refer to occupational and vocational education programs as career education programs.
- Where appropriate, use *students* rather than *children* in order to accurately reflect the age range of those in special education programs.

APPENDIX E SAMPLE FORMS



Page 1

Before test goes to sensitivity reviewer, Test Assembler should fill in all information required above the double line and check all appropriate boxes.

TEST SENSITIVITY REVIEW REPORT FORM

Form Designation ____________

Test Name ____________

Project/Job ____________

Program ____________

Test Assembler ____________

☐ Test assembler requests sensitivity review by subject matter specialist

☐ Special considerations (use in another country, given to handicapped candidates, etc.)

TEST SPECIFICATIONS

I. ☐ Final Form Test Specifications require multicultural material (including minority groups and women)

II. ☐ Final Form Test Specifications do not require multicultural material (but such items are in the test)

III. ☐ This is a pretest for final form with specifications described above

☐ This pretest is required to have multicultural material

☐ This pretest is not required to have multicultural material

Please specify consideration: ____________

Page 2

Test Sensitivity Reviewer ____________ Date received ____________

☐ This is a preliminary review (before editing)

☐ This is a mandatory review (after editing)

OUTCOME OF REVIEW:

☐ Test is approved; no changes required

☐ Changes are recommended (see comments)

☐ Test is acceptable as revised

☐ TD Director consulted on ________ date

☐ Test referred to arbitration on ________ date

Comments: ____________

Required signatures (indicating test is acceptable to both reviewer and assembler as is or as revised)

________ date ____________ Sensitivity Reviewer

________ date ____________ Test Assembler

Total number of items in this test ________

In the boxes below, list the item number in each category. An item referring to more than one subgroup should be listed under each subgroup mentioned in that item; that is, an item mentioning a Black woman should be listed under both female and Black Americans.

		Total
Females		____
Males		____
Asian Americans		____
Black Americans		____
Hispanic Americans		____
Native Americans		____
Others (Specify)		____

Test Review:

☐ Test meets its specifications for inclusion of multicultural material.

☐ Test does not meet its specifications for inclusion of multicultural material.

Item Review:

☐ No comments on items

☐ See comments on items below

☐ No comments on balance

☐ See comments on balance below

Pages 3-4

Comments by Test Sensitivity Reviewer:
Please indicate item number, suggestion, and reason for requesting revision.

Response from Test Assembler:
Please indicate whether revision has been made.
If revision has not been made, please explain why.

If comments continue on attached sheets, please indicate number of pages attached.

Test Sensitivity Reviewer ________ No. of pages

Test Assembler ________ No. of pages

27

How to Fill Out the Blue S.R. Form

Page 1

It is the test assembler's responsibility to see that the entire portion above the double line is filled in before the test goes to a sensitivity reviewer.

Page 2

1. List item numbers on the appropriate lines.
2. List an item in more than one box if applicable. For example, an item about Frederick Douglass would be listed under both "Males" and "Black Americans."
3. No matter how many males and/or females are listed in a single item, enter only the item number. Do not list the number of people in the item. For example, enter only the number of an item based on a chart of ten United States vice-presidents. It is not correct to enter the item as having 10 males.
4. At the bottom of page 2, check the appropriate boxes. If, on page 1, the test assembler has indicated that the test does not require multicultural material, leave the left-hand boxes blank.

Pages 3-4

1. Write your comments on individual items (or sets of items for a passage).
2. Please confine your comments on these pages to sensitivity issues. If you want to raise other issues, make your comments on a separate piece of paper.

Sign Off

1. If the test requires no changes, check the first box on the lefthand side of the first page and sign off on the test. Also, sign off on the appropriate line on the Test Assembler's Control Sheet.
2. If you have made suggestions on revising items or have requested that items be removed, check the second box on the lefthand side of the first page and return the test to the test assembler or appropriate sensitivity review routing coordinator.
3. DO NOT sign either the blue form or the Test Assembler's Control Sheet until the test assembler has responded to your comments.
4. If test assembler's responses are satisfactory, check the third box on the lefthand side of the first page, sign your name on the righthand side, and sign the Test Assembler's Control Sheet.
5. If you and the test assembler cannot agree on changes and/or deletions, check with your divisional sensitivity review coordinator and make arrangements to consult with the TD director. DO NOT sign the blue form or the Test Assembler's Control Sheet.
6. If step 5 does not result in a solution, indicate that the test will go to arbitration. DO NOT sign the blue form or the Test Assembler's Control Sheet.

PUBLICATIONS SENSITIVITY REVIEW FORM

Review Date ____________

Title ____________

Project Director ____________ P J ____________

Sensitivity Reviewer ____________

Results of Review ☐ OK ☐ Revision recommended

Reviewer's Comments: ____________

Project Director's Response: ____________

Approved (Sensitivity Reviewer) ____________

Date ____________

THIS SHEET IS TO BE ATTACHED TO THE SENSITIVITY REVIEW FORM AND REMAIN A PART OF THE PERMANENT RECORD.

Sensitivity Review Arbitration Control Sheet

I have reviewed the

☐ ______________ Test, form ______________

☐ Publication, title: ______________________

and discussed my comments with the assembler/project director. We have been unable to reach a satisfactory resolution of my concerns as explained on the attached sheet(s).

Signed ______________________________
Reviewer

I disagree with the changes as identified with the sensitivity reviewer. We have been unable to reach a satisfactory resolution of my concerns as explained on the attached sheet(s).

Signed ______________________________
Assembler/Project Director

I have been notified of the need for arbitration on the test form/publication designated above.

Signed ______________________________
Sensitivity Review Coordinator

I have discussed the disagreement as described on the attached sheets with the assembler/project director and the reviewer. I am aware that the matter is being sent to arbitration.

Signed ______________________________
TD or Division Director

Having reviewed the written attachments to this control sheet, the arbitration panel has decided as follows: (Continue on a separate sheet if needed.)

Date ______________________ Signed 1. ______________________

2. ______________________

3. ______________________

APPENDIX F

SENSITIVITY–RELATED SECTIONS OF ETS OFFICIAL DOCUMENTS

ETS Standards for Quality and Fairness

Test Development Guideline

Prepare, with appropriate advice and review, specifications for each test that cover the following:

- Sensitivity — requirements for materials reflecting the cultural background and contributions of major population subgroups.

Test Development Guideline

Review individual items, the test as a whole, and descriptive materials to assure that:

- language, symbols, words, phrases, and content that are generally regarded as sexist, racist, or otherwise potentially offensive, inappropriate, or negative toward major subgroups are eliminated.

Accountability Guideline

Review publications and other materials to eliminate language or material generally regarded as sexist, racist, or otherwise offensive or inappropirate.

Memorandum for: COLLEGE BOARD TEST DEVELOPMENT
COPA TEST DEVELOPMENT
SHEP TEST DEVELOPMENT

cc: Ms. Dwyer
Mr. Kimmel
Mr. Klein

info. cc: Test and Publications Editors

Subject: Test Sensitivity Review:
Guidelines for Tallying
Balance

Date: February 26, 1987

From: R. W. Adams
J. Hsia
G. D. Saretzky

Attached are detailed guidelines for tallying items and determining balance for sensitivity reviews of tests. We recommend that these guidelines be placed after Appendix F in the Guidelines and Procedures section of your sensitivity review notebook. Please keep the following points in mind:

1. The tallying and balance guidelines are a part of the overall Guidelines. Even though the focus here is on counting and representation, all other aspects of the review guidelines are to be considered as well during the review process.
2. Although all pretests are to be reviewed to eliminate offensive language, pretests are not required to be balanced in their representation of women or members of minority groups unless the pretest specifications specifically call for such material. (See page 14 of the Guidelines and Procedures.)
3. Test assemblers are urged to inform committees, consultants, collaborators, and survey recipients of the general ETS goals for minority and women's representation and to make certain that all item writers understand the need to meet, when possible, the basic ETS standard for fairness and quality.

/ddmh

Attachments

Test Type: CONTENT

Definition: The CONTENT test is designed primarily to measure knowledge that is specific to a subject; tests measuring knowledge of economics, United States history, literary history, physics, and the like fall into this category. It is expected that such tests will have detailed specifications based on committee recommendations and reflecting, to the extent possible, current curricula.

In general, the items in a CONTENT test ask questions that require the test taker to make use of course-specific information to answer the question posed or to reason through to the correct answer. Examples include, but are not limited to:

> In which of the following circumstances was the National Recovery Act proposed?
>
> Which of the following is part of the Bill of Rights?
>
> Who wrote The Autobiography of Alice B. Toklas?
>
> Which of the following is characteristic of the Gothic novel?

Some items in a CONTENT test may be testing not particular knowledge, but a particular skill such as the ability to interpret data in a form typically used in a discipline. For example, an economics test may require the interpretation of economic data in chart form, or a history test may require the interpretation of a historical or demographic map. Nevertheless, the presence of a small number of skills items in a CONTENT test does not change the classification of the test for the purposes of sensitivity review.

Balance: CONTENT tests meet their own specifications and need not be balanced. They should, however, include in their specifications an indication of the way in which the test can reflect, wherever possible, the contributions of women and minority-group members to the discipline. An economics test, for example, may specify that two items will deal with the impact of women in the labor force and two items with minority-owned businesses. If these items are present in the test, the requirements of the CONTENT specifications have been met and the test is acceptable. It is expected that in some subject-matter areas, for example literary history, there will be considerably more source material available concerning women and minority-group members than there will be in others, for example Latin or Classical History. In the event that the CONTENT test contains items where he/she can be used interchangeably, the test assembler should strive for a balance.

In making preparations for the assembly of a CONTENT test, assemblers are urged to do the following:

1. Inform committees, survey recipients, or other consultants about the ETS standards for fairness and quality, and the implications of these standards for test content.
2. Give item writers clear instructions about the need for representational material and, most importantly, guidance to appropriate source material.
3. Identify all items that are intended to fill women's or minority specifications. This is particularly important for figures whose names may be less well-known or misleading (George Eliot) to non-specialists.
4. Make an effort, where appropriate, to include women and minorities in the distracters of items that are not specifically intended to reflect women's and minority contributions or concerns.
5. Include a copy of the test specifications in the workfolder.

Test Type: SKILLS

Definition: The SKILLS test is designed primarily to measure a particular skill (reading, English usage, mathematical problem solving) that is necessary for academic work but is assumed to be part of the test taker's skills preparation. The subject matter of the stimulus material has no special importance for a SKILLS test. For example, a sentence testing English subject-verb agreement can be about canaries, typewriters, or women novelists, just as a mathematics problem can ask about the height of telephone poles or of basketball players who may be male or female.

There seem to be two fundamental types of SKILLS tests:

1. Tests/sections composed exclusively of discrete items.
2. Test/sections composed of some discrete items and some items linked to the same stimulus (sets).

Separate discussions for balance are given below.

Balance: Discrete items only

In SKILLS tests composed exclusively of discrete items, the representation of males and females in people-related items should be approximately equal. At least 10% of the people-related items should be about minority group members and, whenever possible, more than one minority group should be represented.

It must be recognized that some tests will deviate from these requirements for valid reasons. For example, a test designed exclusively for students in Bermuda may have an entirely different balance, perceived or otherwise, than tests designed for use in the United States or, like TOEFL, designed to test the language abilities of non-native speakers who intend to come to the United States. The coordinator of such a test must document such variations (e.g. Test for use exclusively within a non-United States population), and the test assembler must note the variation on the sensitivity review report forms. Still, in most such tests, a balance of male and female representative items should be the goal.

Balance: Discretes and Sets

For SKILLS tests composed primarily of sets related to a large number of stimulus passages, the people-related passages should, wherever possible, have approximately equal male-female representation. Ideally, there should also be about 10% of the people-related passages that concern various minorities. These criteria should be applied with more flexibility than they are in SKILLS tests composed exclusively of discrete items.

For SKILLS tests composed of a small number of stimulus passages, e.g. 3 as in GMAT or 4 as in GRE, only one of the passages needs to have either women or minority representation.

Test Type: MIXED

Definition: The general definition of a MIXED test is that it primarily measures skills but evaluates those skills by means of clearly defined and established stimulus material within a content area. There seem to be two basic types: 1) curriculum based and 2) occupational.

Curriculum-based

The curriculum based SKILLS test is expected to include stimulus materials that are part of a particular curriculum. For example, a French language test evaluates language ability or interpretive ability by means of specific French texts. Similarly, quantitative ability in Engineering uses engineering-specific stimuli identified as central to the discipline.

Occupational

The occupational SKILLS test primarily measures knowledge and skills required of a particular occupation. Such tests may be intended for an all-male population (the Professional Golf Association, for example) or may deal only with women (Obstetrics and Gynecology, for example). The situations and content to which the skills and knowledge are applied thus represent those that are relevant to the occupation.

Balance: For mixed tests the sensitivity review requirements for balance must be used judiciously. The ETS standards must be kept in mind, but the unique orientation of the test must be recognized as well. In assembling such tests, the assembler must plan well ahead to make certain that whatever possibilities exist for representation of women and minority groups are identified and developed as far as the important subject-matter of the test will allow.

Curriculum-based

A curriculum-based MIXED test will generally conform to whatever domain has been recommended by external committees, consultants, or internal specialists. Although this domain may have its own requirements (French language and literature, 20th century American literature), it nevertheless remains the responsibility of the test assembler to ensure that, whenever reasonable and possible, representative material is developed. Thus, whereas a CONTENT test may specify "two contemporary Black women writers," the MIXED test with only a specification for "ten 20th century American items" should also contain some representation of American women and American minorities in either stimulus material or individual items. Obviously there are limitations, depending on the subject area; the test assembler is only asked to exhaust whatever possibilities exist for achieving representation of women and minority groups consistent with the general guidelines.

There will be curriculum-based MIXED tests that combine significant numbers of skills items and significant numbers of content items in sections that may or may not be separately timed. For example, a test on Spanish language and culture might have 60% Spanish grammar and vocabulary items (SKILLS because the subject matter of the items could be anything) and 40% Spanish history and culture items (CONTENT because the items measure knowledge of a specific subject). In this type of MIXED test, the skills material should be evaluated in a way generally consistent with the balance requirements for SKILLS tests. For example, if there are discrete items dealing with grammar and vocabulary, those referring to people should be approximately equal in male-female references. Similarly, the content material should be evaluated in a way generally consistent with CONTENT tests; the assembler should provide detailed specifications for such content portions of the test.

The sensitivity reviewer must remember that MIXED tests of this nature are likely to have a clear content base and general "culture" orientation. This means, therefore, that Spanish tests, for example, would be expected to have some representation of Hispanic-Americans but would not necessarily be expected to have Black American, Asian/Pacific American, or Native American/American Indian representation in either the skills or content portions of the test. Similarly, a test like the Bermuda test, designed for a non-White and non-United States population, might be relatively free of any "minority" representation at all.

Occupational

An occupational MIXED test will generally conform to the activities, knowledge, and skills required by an occupation and identified as central by consultants, committees, or internal specialists. Here, as with curriculum-based tests, the test assembler is asked only to make certain that whatever reasonable possibilities for women's or minority representation exist are used.

Guidelines for Tallying

The basic rules for tallying, regardless of the type of test, are as follows:

Items

1. For discrete items, a reference to a male, a female, or a member of a minority group in the stimulus, stem, or options means that the item should be tallied under that group.
2. For discrete items, a reference to both a male and a female, or to both a female and a member of a minority group, means that the item should be tallied once under each of the groups mentioned.
3. For discrete items, a reference to a Black woman should be tallied once for the Black category and once for the female category; similarly a reference to a Hispanic man is tallied once in both categories, and so on for other groups.
4. For discrete items, a reference within one item to several men, several women, or several members of a minority group should be tallied only once for each group.
5. Only United States minority groups should be tallied. For example, a passage about the Japanese writer Mishima Yukio should not be tallied as "Asian American" and a passage about the Maya should not be tallied as "American Indian." Such material can, for completeness, be listed under "other," but cannot be considered in determining whether specifications for balance have been met.
6. References to male or female animals, birds, or mythological creatures should not be counted for balance, and the specific behavioral characteristics of such figures should not be extended to human males or females for any reason. Thus, the behavior of a female bird or a male bear is in no way representative of or stereotypical of human behavior.

Passages

1) For a stimulus or passage with a set of items attached, the sensitivity reviewer should determine whether the stimulus/passage is about men, or women, or a minority group. If it is and the passage is in a SKILLS test, all of the attached items should be tallied for that group.

   In a CONTENT or MIXED test, correct identification of what the passage/stimulus is about is also important. It is best in CONTENT or MIXED tests, however, to maintain a double count—one count for passages and one count for items—since the items may or may not have relevant references and should be counted individually.

2) If a passage/stimulus incidentally mentions a man/woman/minority group but is not about that man/woman/minority group, the reference can be entered once in the item tally (regardless of how many times the name of the man/woman/minority group appears) in the appropriate place on the review form. For example, if the name Martin Luther King Jr. appears twice in a passage about non-violent resistance, the item tally should have one mark for Black and one mark for Male, but the passage should not be considered male-oriented.

   It should be noted that, in a passage about Martin Luther King Jr., once that passage has been categorized the tally does not increase each time King's name appears in the passage.

3) In a CONTENT or MIXED test, passages or works of art that, even though not identified as such, are by women or members of minority groups must be counted in the tally. Thus, a painting by Mary Cassatt or an excerpt from a poem by Langston Hughes, even if no explicit reference is made to women or minorities and even if the artist is not identified in the items, must be counted for the purposes of balance.

   Test assemblers should always identify artists or authors who are women or members of a minority group on cards or flimsies.

4) In a CONTENT or MIXED test, a passage that, for context/subject-matter considerations, must contain "generic he" references or personified objects, should be classified with care. For example, if "the West Wind" is personified as "he" in a poem about the West Wind, the general orientation might be considered male but each individual "he" should not be part of the tally. A single reference to a male West Wind in a poem about summer would not, however, make the poem "male oriented" and a single pronoun probably should not be tallied. In a passage with "generic he" references, the passage should be classified as male oriented, but each he should not be tallied. The items for such a passage will be tallied according to their individual characteristics.

The ETS Sensitivity Review Process: A Commentary for Test Assemblers

Table of Contents

Introduction 3
Inappropriate Language 3
Inappropriate Subject Matter or Tone 4
Inappropriate Underlying Assumptions 7
Stereotyping 8
Lack of Balance 9
Juxtaposition 11
Judging the Items 11

I was breezing along through a chapter on the American Revolution when I did a double take on one sentence. It was as if somebody had stuck a foot out there on the page and tripped my mind as it went by. I looked again, and this sentence jumped out at me: Despite the hardships they suffered, most slaves enjoyed a higher standard of living and a better life in America than in their primitive African homeland. As far as I can remember, this was the first time I was ever enraged.

—Bill Russell*

---

* Russell, Bill, and Branch, Taylor. Second Wind. New York: Random House, 1979.

Introduction

Most of the items and tests the test sensitivity reviewer judges need no change to meet the ETS sensitivity guidelines. However, the reviewer must be aware that some items may be flawed in terms of sensitivity issues. These flaws fall into five categories:

(1) Inappropriate language
(2) Inappropriate subject matter or tone
(3) Inappropriate underlying assumptions
(4) Stereotyping
(5) Lack of balance

The items and passages included in this section have been chosen to illustrate these basic flaws—or the lack of them. Some of these items and passages appeared in tests produced before the ETS test sensitivity review process went into effect; some of them never appeared in tests at all, but were removed from the pool of available items and offered by test assemblers for use in training sensitivity reviewers. Comments following each of the items or passages are intended both to direct attention to the problems sensitivity reviewers found with the material and to help define what is and is not acceptable under the ETS sensitivity review guidelines.

(1) INAPPROPRIATE LANGUAGE

Example

Owing to her detailed and perceptive study of the modern female, Germaine Greer has become a recognized spokesman of the women's liberation movement.

Begin with Her detailed.

(A) having made
(B) has made
(C) made of
(D) became
(E) is becoming

Commentary

Spokesman is not the term that should be used here. *Advocate* or *leader* would be an acceptable substitute.

Example

In an inner city, retail sales and employment are declining and low-income and minority households are increasing. Which of the following policies would be COUNTERPRODUCTIVE to restoring its economic health and vitality?

(A) Approval of a new circumferential freeway outside of the city
(B) Adoption of a regional tax-sharing plan for all new industrial and commercial development
(C) Concentration of federal grant monies for rent supplements to inner-city areas
(D) Diversion of highway trust monies to public transit improvements

Commentary

The term *inner city* carries a great many connotations not necessary to the item. *Downtown area* would avoid those connotations. The inclusion of *minority households* in the stem is irrelevant; the pertinent information is contained in *low-income households*.

Example

Rosa Martinez's vituperative review of the film cast doubt on her ability to assess the worth of cinematic works because that film has been an overnight box office success.

Commentary

In this sentence testing English usage, the designation of a Hispanic woman as a literary critic was undoubtedly meant to show respect for both Hispanics and women. Unfortunately, *vituperative* and the implication that the critic's judgment is valueless make the item affectively negative.

Example

To deal with the problems raised by the women's liberation movement, it demands basic changes in our assumptions about the organization of society.

(A) it demands basic changes
(B) basic changes are what it demands
(C) there are basic changes demanded
(D) people must make the basic changes
(E) we must make basic changes

Commentary

Too often items dealing with minorities and women use the word *problems*, implying that the quest for civil rights or job opportunity brings nothing but trouble and annoyance to the rest of the world. It would be best if ETS items avoided giving support to such a negative view of the changes brought about by the civil rights and the women's movements. The phrase *problems raised* could be changed to *opportunities* or *challenges presented*.

Example

Experience has shown that 75 percent of those hired for a certain job prove to be successful. A test is administered to 80 applicants and the 40 men with the highest test scores are hired. If it turns out that the test has zero validity, what percent of these men should be expected to be successful?

(A) 0% (B) 40% (C) 50% (D) 75% (E) 80%

Commentary

The use of *men* in this item is unnecessary, confusing, and in violation of the ETS guidelines. A neutral word like *applicants* or *test takers* can easily replace *men*.

(2) INAPPROPRIATE SUBJECT MATTER OR TONE

Example

Just as the——of a new species of insects is certain to have a profound effect on the——of a river valley, so a large immigration of a new race or class is bound to destroy the social equilibrium of a city.

(A) exodus. .topography
(B) influx. .ecology
(C) mutation. .geology
(D) discovery. .population
(E) extermination. .stability

Commentary

The sentence implies that immigrants of a different race or class are, like destructive insects, bound to destroy the territory they enter. These suggestions are totally inappropriate in an ETS test. The sentence is affectively negative.

Example

Both candidates agreed that such minorities must be given an opportunity to advance, to seek justice, and to _________ the kind of special treatment that might make up in part for past inequities.

Commentary

This sentence (testing English usage) implies that minorities are passive, only awaiting the paternalism of the majority to improve tneir lot in life. The sentence might be revised as follows:

Both minority candidates agreed that minority people must take this opportunity to speak out, to seek justice, and to _________ the kind of education that might enable them to make up, in part, for past inequities in employment.

Example

People have been in the Americas for more than 38,000 years. Whites have been around for less than five hundred. It is presumptuous for anyone to pretend that the Chicano, the "Mexican-American," is only one more in the long line of hyphenated immigrants to the New World. I reject the semantic games of the sociologists who identify us as Mexican-Americans. Our insistence on calling ourselves Chicanos stems from a realization that we are not just one more minority group in the United States.

We are, to begin with, a powerful blend of indigenous America with European-Arabian Spain. During the three hundred years of New Spain, only 300,000 whites settled in the New World, and most of these were men. There were so few white people at first, that ten years after the conquest in 1531 there were more black men in Mexico than white. Africans were brought in as slaves and soon intermarried. Miscegenation went joyously wild, creating many hues, shapes, and sizes, but the predominant strain remained Indio.

Then in the twilight of the conquistadores' domination of New Spain, the Indios suffered the fate of a colonized people. Rejected by the Spanish father, they clung to their Indian mother and shared her overwhelming sense of loss. The revolution of the thirteen colonies of New England did not touch us, the descendants of the Indios, until half a century later. Having formed a nation, the colonies eventually looked south for their own conquests and decided to "liberate" Texas from Mexico. Mexico itself was bleeding from internal conflict and was ill-equipped to defend its people in the war that made Texas part of the new country. Amid this so-called liberation, the American Indio remained forgotten.

Who then are the residents of the United States known by the Chicano as Anglos? They are transplanted Europeans, with pretensions of native origins. Their most patriotic cry is basically the retort of one immigrant to another. Feeling truly American only when they are no longer the latest foreigners, they brandish their Americanism by threatening the new arrival: If you don't like it here, go back where you came from.

Now the Anglo is trying to impose the immigrant complex on the Chicano, pretending that the "Mexican-Americans" are the most recent arrival. But we will not be deceived. In the final analysis, frijoles, tortillas, and chili are more American than the hamburger. We do not suffer from the immigrant complex. We left no teeming shores in Europe, impatient and eager to arrive in New York. No Statue of Liberty ever greeted our arrival in this country. We did not, in fact, come to the United States at all. The United States came to us.

Commentary

This reading passage was rejected primarily because of its inflammatory tone, which might be upsetting to various groups of test takers. The material within it is controversial and affectively negative; in this case, the material is potentially offensive to members of both the majority and minority groups.

Example

The only Oriental boy in a class of five-year-olds always looks down when the teacher addresses him. Of the following, the most reasonable assumption the teacher can make about his behavior is that he

(A) probably feels guilty and thinks he has done something wrong
(B) has learned to lower his eyes for a particular reason
(C) may have trouble with his vision and should have his eyes checked
(D) does not pay attention when spoken to because he is thinking of other things
(E) may be emotionally disturbed and should be observed by the school psychologist

Commentary

The subject matter of the item is appropriate for a test given to teachers, who should be aware of different cultural traditions among their pupils. The difficulty with the item lies in the vague key, option B, and the affectively negative options, each of which, when placed with the stem, is demeaning or insulting. The test taker who answers this incorrectly and who does not know that the option chosen is incorrect may well have negative ideas reinforced.

The ethnocentric word *Oriental* should be changed to the preferred designation, *Asian-American*. The item can be revised in several ways to avoid the negative qualities of the options. For example:

Which of the following is the most probable reason why the boy looks down?

(A) To concentrate better
(B) To show respect
(C) To avoid embarrassing the teacher
(D) To ask permission to question the teacher
(E) To avoid showing disagreement with the teacher's remarks

Example

Frequently there is a time lag between the statement of a managerial policy and the implementation of that policy. This appears to be particularly true with regard to the acceptance of women in management positions. According to our survey findings, women interested in management or professional careers still face social and psychological barriers, despite recent changes in policies on the employment of women.

The responses we received to the case examples reflect two general patterns of sex discrimination: (1) There is greater organizational concern for the careers of men than there is for those of women, and (2) There is a degree of skepticism about women's abilities to balance work and family demands. Underlying these patterns of discrimination there is an assumption that is not at first apparent from the survey findings: it appears that women are expected to change to satisfy the organization's demands. For example, written comments from participating managers often suggest that women must become more assertive and independent before they can succeed in some of the situations described in the case examples in the survey. These managers do not see the organization as having any obligation to alter its attitudes toward women. Neither, apparently, are organizations about to change their expectations of men. Perhaps because it is expected that the job will eventually "win out" over the family, a man is given the time and opportunity to resolve conflicts between home and job. This in itself says a great deal about how organizations might conceive of a man's relationship with his family.

Another conclusion we can draw is that when information is scant and the situation ambiguous, managers tend to fall back on traditional concepts of male and female roles. Only when there are clear rules and qualifications do both women and men stand a chance of breaking out of the stereotyped parts usually reserved for them.

When the results of this survey are extrapolated to the total population of American managers, even a small bias against women could represent a great many unintentional discriminatory acts that potentially affect thousands of career women. The end result of these various forms of bias might be great personal damage for individuals and costly underutilization of human resources. If managers are sincere in wanting to encourage all employees equally, they ought to examine their own organizations' implicit expectations of both men and women to see whether these expectations reflect some of the same traditional notions revealed by the survey. Identification of these biases would help managers to move toward the goal of equal employment opportunity for all.

Commentary

This passage was undoubtedly chosen as appropriate for showing women's concerns in a test meant for applicants to graduate schools of business. Unfortunately, the passage does not present a positive picture of what women entering business management can expect and can therefore be considered affectively negative for women taking that test. It might not be considered inappropriate in another context, however. In any case, it would be better to use the term "women" rather than "career women."

(3) INAPPROPRIATE UNDERLYING ASSUMPTIONS

Example

In order to work effectively with members of a minority group, the most important consideration is for the social worker to

(A) be aware of his or her own self, values, and biases
(B) study the language of the minority group
(C) be sympathetic and nondiscriminating
(D) live among or close to the minority-group members

Commentary

The item assumes that the social worker is not a member of a minority group. This underlying assumption is invalid, and some of the options are also patronizing.

Example

The fact that black community organizations perceive that economic development meets their needs does not by itself justify a federal investment in economic development programs. There are, however, at least two important programmatic reasons for establishing economic development programs with broad-based community support. First of all, it is becoming increasingly more difficult for any federal program to operate in a black community without reference to the social and political forces within the community. The civil rights movement and several years of operating community action programs have made a change in the environment. Black people have become more skilled in the techniques of organization and of communication with the white community. As a result, it has become virtually impossible to implement any meaningful program without active community participation.

The second program reason for community control is directly related to the fact that the social utility of economic development involves multiple benefits. As long as programs involve single, separate, quantifiable outputs such as total employment, total number of houses built, etc., a strong case can be made for having the ultimate control of the program in the hands of the technicians who are better equipped to achieve these goals and to optimize the various combinations of cost-benefit relationships. However, community economic development requires that trade-off decisions be made involving nonquantifiable comparisons. Given the fact that the state of the art of cost-benefit analysis is, and in the near future will continue to be, much too crude to permit any semblance of objective cross comparisons of social benefits, the question becomes, "Who should decide between social benefits?" If someone has to make these judgments, it is reasonable to assume that the perception of the community that has to suffer any mistakes is a better guide than the perception of outside professionals who lack both the conceptual framework and the data for rational analysis.

This does not mean that residents can develop their community without outside help. This is particularly true in programs of business development that by nature involve complex interrelationships between people, require considerable technical competence, and presume a certain common frame of reference among participants. The trick, therefore, is to find the combination of community control and technical capability that will produce responsive policies and competent programs. In effect, there would be a partnership between black institutions and the establishment, with government equipping the institutions with the fiscal base for negotiation.

7

Commentary

This passage basically sets up a "we-them" situation, with the "we," who are knowledgeable and technically skillful, proposing to help the "them," who cannot provide any of the knowledge or skills required out of their own resources. The author is not unaware of the need for community involvement and cooperation or the skills that the community will be able to provide. However, the author assumes that the entire community will be ignorant of the knowledge required and deficient in the skills needed. This underlying assumption is what leads to the affectively negative aspects of the passage.

Statements like "Black people have become more skilled in the techniques of . . . communicating with the white community" reinforce the inappropriate underlying assumption. It would be more appropriate to suggest that Black and White communities have become more skilled in communicating with each other rather than implying that any failures in communication are the responsibility of the Black community alone.

(4) STEREOTYPING

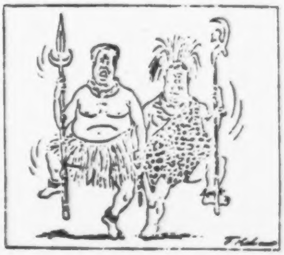
Example

Khrushchev's gift to history is, and always was, himself. Khrushchev's greatest qualities, those that distinguished him from all other Soviet leaders, were his energy, his enthusiasm, his confidence in himself and in others. It was his prodigal personality, his ability to confess a mistake and reverse himself, his explosive unpredictability, that did more than anything else to spring the genie of spontaneity out of the bottle of repression in which Stalin had contained the Russian spirit for 30 years. Khrushchev was the quintessential Russian peasant. He was cunning and sly. He was given to the charming, fantastical Russian kind of lying called *vranyo* and to extremes, like the *muzhik* who works hard and then spends days on a drinking spree. Coming at the moment of Russian history when he did, Khrushchev's great contribution was his confidence in the Russian people and his effort to give them confidence in themselves.

Commentary

The stereotype in this passage is of Russian peasants, a group not explicitly covered in the guidelines. However, the passage does present a stereotype; the stereotype is offensive, and the guidelines do indicate that offensive stereotypes are to be avoided. The material could be affectively negative for certain members of the population taking the test.

Example



The cartoon above from the early 1960s depicts

(A) a newly revived tribal dance
(B) communist anger at American involvement in Vietnam
(C) the displeasure of communist leaders over the closing of the Suez Canal
(D) efforts to increase communist influence in Africa
(E) the resentment of Mao and Khrushchev at African attempts to mediate the Israeli-Arab conflict

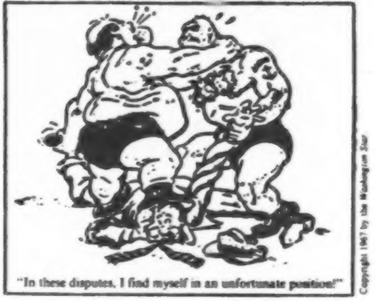
Commentary

The cartoon is offensive because it stereotypes Africans as primitive, spear-carrying people in grass skirts or leopard skins. The figures are meant to represent Khrushchev and Mao.

Although material of a satiric nature often raises issues for sensitivity reviewers, it is possible to use cartoons and the like that meet ETS sensitivity review guidelines. The following material is acceptable:

Example

Questions 46-48 are about the following cartoon.



"In these disputes, I find myself in an unfortunate position!"

Copyright 1967 by the Washington Star

46. The man on the bottom in this cartoon represents
(A) the federal government
(B) a labor mediator
(C) the consumer
(D) the farmer

47. What is the main point of the cartoon?
(A) Labor-management disagreements often lead to violence.
(B) The government has no power to stop strikes.
(C) Farmers have little influence on national politics.
(D) The public is often hurt by labor-management disagreements.

48. The way the fighters are drawn suggests the artist believes that
(A) both labor and management obey the rules in their disagreements
(B) both labor and management are powerful forces
(C) the government has too much control over labor and industry
(D) both labor and management want help in solving their disagreements

Commentary

The cartoon and the questions following it are acceptable. The exaggeration in the cartoon and the choices in the items do not present a derogatory picture of any group.

(5) LACK OF BALANCE

Balance in a test, for the purpose of sensitivity review, generally involves including items that present males and females in approximately equal numbers, showing women as well as men as active participants in the world at large, and presenting the contributions of members of various minority groups or describing the history and culture of such groups as well as those of the majority group.

The following group of items illustrates another kind of balance that the sensitivity reviewer may comment on. Further, it is important to note that the items appeared in a descriptive booklet (all ETS publications are subject to a mandatory sensitivity review) and that the question raised for the sensitivity review is not covered directly in the ETS sensitivity review guidelines. The items are being used to describe the kinds of items that appear in a humanities test.

Careful consideration of balance is most appropriate in operational sections/tests made up of discrete items. In a reading comprehension pretest made up of only two or three passages, the test assembler may have content specifications that do not include minority/female representation. Such pretests cannot be challenged for balance.

Examples

1. Often read as a children's classic, it is in reality a scathing indictment of human meanness and greed. In its four books, the Lilliputians are deranged, the Yahoos obscene.

   The passage above discusses

   (A) *Tom Jones* (B) *David Copperfield* (C) *The Pilgrim's Progress* (D) *Gulliver's Travels*
   (E) *Alice in Wonderland*

2. Which of the following deals with the bigotry an anguished Black family faces when it attempts to move into in all-White suburb?

   (A) O'Neill's *Desire Under the Elms*
   (B) Miller's *Death of a Salesman*
   (C) Williams' *A Streetcar Named Desire*
   (D) Albee's *Who's Afraid of Virginia Woolf?*
   (E) Hansberry's *A Raisin in the Sun*

3. Which of the following has as its central theme the idea that wars are mass insanity and that armies are madhouses?

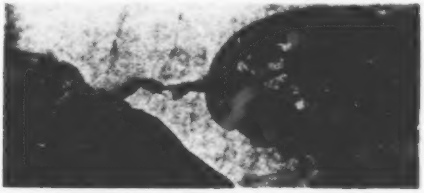
   (A) *Catch-22* (B) *Portnoy's Complaint* (C) *Lord of the Flies* (D) *Heart of Darkness*
   (E) *Vanity Fair*

4. Which of the following is often a symbol of new life arising from death?

   (A) A gorgon (B) The minotaur (C) A unicorn (D) A griffin (E) The phoenix

5. Which of the following musical forms is divided into the sections: Kyrie, Gloria, Credo, Sanctus, Benedictus, Agnus Dei?

   (A) A symphony (B) A piano concerto (C) A mass (D) A madrigal (E) A cantata



1. The work pictured above is

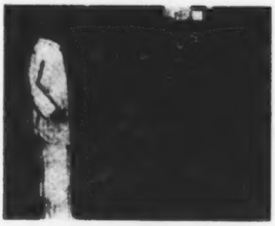
   (A) a fresco (B) a stabile (C) a woodcut (D) an illumination (E) an etching

2. The theme of the work is the

   (A) sacrifice of Isaac (B) expulsion from Eden (C) reincarnation of Vishnu
   (D) creation of Adam (E) flight of Icarus

3. The work is located in the

   (A) Alhambra (B) Sistine Chapel (C) Parthenon (D) palace at Versailles
   (E) Cathedral of Notre Dame



This painting is a visual allusion to which of the following pictorial themes?

(A) The Annunciation (B) The Flight into Egypt (C) The Adoration of the Magi
(D) The Pietà (E) The Descent from the Cross

Commentary

The question of balance raised by these items is whether only those familiar with Christian tradition can achieve a good score on the test. Christianity has indeed influenced the music, painting, and other arts of Western civilization, and testing knowledge of these influences may certainly be appropriate, depending on the purposes of the test. The questions the assembler and the test sensitivity reviewer must decide are:

- Do these items indeed reflect the proportion of questions on the test dealing with Christian tradition?
- If they do, is such emphasis on detailed knowledge of Christian tradition justifiable?

JUXTAPOSITION

Sometimes two items are acceptable from the sensitivity reviewer's point of view, but they present a problem because they are juxtaposed. Juxtaposition can permit an unwelcome and unintended association between ideas. For example, an item dealing with Black women followed by one dealing with welfare might cause some test takers to make an unwarranted association of Black women with welfare recipients. Such items should be separated.

JUDGING THE ITEMS

Nothing involving human relationships is ever cut-and-dried, and reviewing materials for potential offensiveness to particular groups of people is no exception. The guidelines for reviewing are just that—guidelines. They do not indicate exactly how every item or passage undergoing sensitivity review is to be interpreted or under what circumstances material that might be regarded as inappropriate for one test becomes acceptable—or at least tolerable—for another. Because there is leeway for debate about some items and their use in a particular test, sensitivity reviewers are encouraged to discuss with other sensitivity reviewers material that they consider potentially offensive.

The need for discussion is particularly apparent when the sensitivity reviewer considers the material potentially offensive enough to be removed from a test because no amount of rewording will make the material acceptable. Before the sensitivity reviewer embarks on such a course, it is important that he or she determine from discussions with other sensitivity reviewers whether the recommendation to remove material from a test is an idiosyncratic or individual response or the informed response of a group of sensitivity reviewers. Throughout the process, discussion of disputed items is encouraged not only among sensitivity reviewers but also between the sensitivity reviewer and the assembler and among all the parties interested in the outcome of the dispute.

Example

My grandmother's notorious pugnacity did not confine itself to the exercise of authority over the neighborhood. There was also the defense of her house and her furniture against the imagined encroachments of visitors. With my grandmother, this was not the gentle and tremulous protectiveness of certain chronically frail people, who infer the fragility of all things from the brittleness of their own bones and hear the crash of mortality in the perilous tinkling of a tea cup. No, my grandmother's sentiment was more autocratic: She hated having her chairs sat in or her lawns stepped on or the water turned on in her sinks, for no reason but pure administrative efficiency; she even grudged the mailman his daily promenade up her sidewalk. Her home was a center of power, and she would not allow it to be insulted by easy or democratic usage. Under her jealous eye, its social properties had withered and it functioned in the family structure simply as a political headquarters. Family conferences were held there, consultations with the doctor and the clergy; unruly grandchildren were brought there for a lecture or an interval of thought-taking; wills were read and loans negotiated. The family had no friends, and entertaining was held to be a foolish and unnecessary courtesy required only by the bonds of a blood relationship. Holiday dinners fell, as a duty, on the lesser members of the organization: Sons and daughters and cousins respectfully offered up Baked Alaska on a platter, while my grandparents sat enthroned at the table, and only their digestive processes acknowledged the festal nature of the day.

Commentary

Some test sensitivity reviewers thought this passage inappropriate because it describes an unpleasant woman. Others had no objection to the passage, because it was obviously describing an individual woman and not all grandmothers.

Presented to a panel of experts in literature assembled at ETS to discuss sensitivity issues in the testing of literature, the passage was approved. The crucial factor for the panel was that the person described in the passage is obviously a character of considerable individuality and not stereotypical in any way.

Example

The Mescalero Apache tribe is one of seven linguistically and culturally related peoples whose aboriginal territories stretched over large sections of present-day southwestern United States and northeastern Mexico. The Mescalero were characterized by an economic system that harmonized well with their challenging environment. In late historic times they attempted desultory farming along some watercourses, but the severe weather and short growing season of the mountains and the precarious water supply of the lowlands did not encourage cultivation of the soil. Thus the Mescalero were forced to depend on hunting and the gathering of wild harvests.

Such an economy required mobility; there had to be readiness to follow the food harvests when and where they matured and to move from one hunting area to another when the supply of game dwindled. A concentration of population was inappropriate to such techniques of food procurement. As a result, the population was thinly dispersed over the immense range.

Since most economic errands were carried out in small groups, there was little incentive for highly centralized leadership. It is probable that never in its history did the tribe have a single leader who was recognized and followed by all. Rather, the Mescalero leader, or "chief" (literally "he who speaks"), was, as his title suggests, a respected adviser drawn from the heads of the families who tended to camp and move together.

Since he had no coercive power, he had to understand what his followers were willing to do. Serious misjudgments or unpopular counsel might cost him his position or a portion of his followers. Theoretically, the office of the leader was not hereditary; in practice, there was a tendency for sons of leaders to succeed their fathers. This was informal, however, not absolute. Typical situations which required a leader's judgment included such problems as whether to move to another site because of poor luck in hunting, repeated deaths, epidemic disease, or the proximity of enemies; whether to sanction a raid or war party; whether to sponsor an important social or ritual event to which outsiders might be invited; and what to do about disruptive behavior such as the practice of witchcraft. The ability to lead successful raids and war parties, as well as to sanction them, was a great asset for a leader; such expeditions meant booty, and this made it possible to distribute favors widely. In a society where generosity was one of the cardinal virtues, such activity built and sustained the good will so important to a leader.

Commentary

The main objection to this passage, in discussions among sensitivity reviewers, was that it makes no mention of women. Further, some of the language was considered demeaning, for example, the phrase "desultory farming." Others who reviewed the passage had no objection either to the failure to mention women or to the language of the passage. They held that the society being depicted was a male-dominated society, that women in that society were subjugated, and that it might be more insulting to women to describe that subjugation than to omit mention of women entirely. They had no criticism of the use of the word *desultory*, considering that a nomadic people would not farm in any other way. The basic argument of these reviewers was that the Mescalero Apaches lived a life different from that of modern inhabitants of North America and that failure to describe that life accurately was, in a sense, an admission that that way of life was to be regarded as not so much just different but inferior. The counterargument was that not all readers of the passage would be knowledgeable about various kinds of societies and would tend to view as inferior a society quite different from what they considered the best or the norm.

A panel of historians was invited to ETS to discuss with staff various issues that had caused concern among both test assemblers and sensitivity reviewers. In its review of the materials presented to it, this panel made a clear distinction between material that it considered suitable for history tests and material that would be appropriate in reading passages. The view of panel members was that no subject of importance is to be avoided in a test designed to be taken by students of history, although care should be used in the presentation of material. Merely including the date or source of an opinion would suffice for some material. They stated that history was not always pleasant and that unpleasant aspects of history should be studied and knowledge of those areas of history should be tested. They also maintained, however, that considerable circumspection was needed in choosing passages about minority groups and the history of minority groups for use in reading tests. Unlike history students, takers of such tests cannot be expected to supply or understand a context for a particular idea that might be potentially offensive or disturbing.

The historian who discussed the passage about the Mescalero Apaches with ETS staff, an American Indian himself, asserted that descriptions of members of the various Native American nations and tribes in tests and other materials contain three major faults:

(1) Native Americans are dealt with as peoples of the past. Very little attention is paid to the American Indians living, working, accomplishing today.

(2) They are defined in two ways only—as lovers of nature or as fierce warriors. (He could not decide, he said, which stereotype he disliked more.)

(3) In most materials, American Indians appear to have lived in a society without women.

Given these stereotypes, the passage on the Mescalero Apaches is to be considered inappropriate for use in a test.

Example

George Bernard Shaw explicitly advises women to be selfish. Of course, his play *Major Barbara* reminds us that selfishness is not for females only. In Undershaft, the munitions millionaire, selfishness is bolder in outline and broader in scope than anything Dorothea or Ibsen's Nora or Undershaft's daughter could achieve. Undershaft will see the world blown apart by his munitions before he will submit himself to the degradations of poverty. None of the women quite reaches this pinnacle of assertion. Yet the actual pattern is not different; and however small the framework, however delicate the tracing, the quality of selfishness in women needs to be emphasized just because it is so difficult to achieve. That remark may sound dubious to readers who know selfish women in life and literature; but these are examples of petty selfishness, not grand selfishness, and of the old-fashioned, not the Shavian new woman.

The grand selfishness Shaw recommends is not self- serving, but self-respecting; it does not result in petty self-seeking, but in a rehabilitation of the idea of the self. Selfishness is the opposite of meekness, humility, and self-sacrifice (the so-called womanly virtues), not the opposite of generosity and altruism (the virtues of strength). In a badly arranged world, meekness and acquiescence are dangerous virtues.

In fact, two pieces of spiritual advice sum up this little book and could equally well stand for the whole of Shaw's advice to women. The first is the Johnsonian, "My dear friend, clear your mind of cant." The second, a comment with reverberations, is: "Always have the highest respect for yourself, and you will be too proud to act badly."

Whether this quality is called self-respect, pride, or egotism, it has always been most difficult for men to accept in women or for women to accept in themselves. According to Lionel Trilling, it is this quality in the heroine of Jane Austen's *Emma* that dismays the critics and introduces an equivocal note into their judgments of the novel. It is this that is distasteful in the heroine of Charlotte Bronte's *Shirley*, who is a plausible ancestor of Shaw's Lydia Carew: a self-willed, egotistical woman who defies the world to censure her, who feels sure that "what everybody knows" is wrong and that her own unconventional views are right. One measure of the change in social atmosphere between 1849, when *Shirley* was published, and 1886, when Shaw completed *Cashel Byron's Profession*, is Shirley's assertion of her right to certain masculine prerogatives, which she self-consciously pursues as an end in itself. Lydia, on the other hand, treats her feminine self-assertion quite absentmindedly, with the main force of her attention focused on the objects to which this assertion admits her: interesting studies and rationalized rules of behavior.

39. According to the author, the actual existence of selfish women may lead some people to conclude that
    - (A) all women are selfish
    - (B) women who are selfish are simply acting normally
    - (C) society does not demand either selfishness or selflessness of women
    - (D) self-sacrifice is only one of the possible patterns of action available to women
    - (E) it is easy for women to exhibit selfishness

40. The author uses the example of Undershaft to
    - (A) show Shaw's tendency to exaggerate
    - (B) contrast Shaw's characters with Ibsen's
    - (C) demonstrate the realism of Shaw's characterizations
    - (D) present a contrasting model for Shaw's work
    - (E) rebut Shaw's primary contention

41. The author meets an anticipated objection to the statement that it is hard for women to behave selfishly by
    - (A) presenting an example
    - (B) making a distinction
    - (C) citing an authority
    - (D) describing historical conditions
    - (E) examining literary attitudes

42. Which of the following provided the clearest instance of the selfishness that Shaw recommends?
    - (A) A woman who allows others to work to support her but does not help to provide for her family's needs
    - (B) A woman who makes other members of her household defer to her preferences
    - (C) A woman who insists on training for and practicing a profession that she chooses
    - (D) A woman who marries for social advance and does not attempt to make her husband happy
    - (E) A woman who obtains the largest part of an inheritance by flattering a wealthy grandparent

43. As described by Lionel Trilling, the response of critics to Jane Austen's *Emma* was influenced by the fact that they
    - (A) found it hard to accept the character of an assertive woman
    - (B) did not understand the tradition in which the novel stood
    - (C) failed to appreciate the subtlety of Jane Austen's characterization
    - (D) thought it unsuitable for women to write novels
    - (E) found the novel to be ambiguous in its values

44. It can be inferred that the qualities of Charlotte Bronte's heroine in *Shirley* are "distasteful" to
    - (A) the author of the passage
    - (B) many feminists
    - (C) many readers of the novel
    - (D) George Bernard Shaw
    - (E) Lionel Trilling

45. It can be inferred that Lydia Carew is which of the following?

(A) A nineteenth-century novelist
(B) A career woman in Shaw's *Major Barbara*
(C) A political figure
(D) A member of a group that agitated for women's rights
(E) A character in *Cashel Byron's Profession*

46. It can be inferred that the author views the lessening, between the times of Shirley and Lydia, in the self-consciousness of women who asserted themselves as

(A) a cause of deterioration in sexual relationships
(B) an adoption of ladylike behavior
(C) a setback for society
(D) progress made by women
(E) a result of women's renunciation of egotistic behavior

Commentary

The consensus of sensitivity reviewers who looked at this passage was that they would approve its use in a test. They also recognized that some sensitivity reviewers might object to the passage and to items 39 and 42 particularly. However, the group of reviewers approving the passage deemed that it does not depict women either negatively or stereotypically and that items 39 and 42 were acceptable in that each was being used to define selfishness, a crucial point in understanding the passage and the attitude presented.

Example

The days between Christmas Day and New Year's were allowed the slaves as holidays. During these days all regular work was suspended, and there was nothing to do but keep fires and look after the stock. We regarded this time as our own by the grace of our masters, and we therefore used it or abused it as we pleased. The holidays were variously spent. The sober, industrious ones would employ themselves in manufacturing corn-brooms, mats, horse-collars, and baskets, and some of these were very well made. Another class spent their time in hunting opossums, coons, rabbits, and other game. But the majority spent the holidays in sports, ball-playing, wrestling, boxing, running, foot-races, dancing, and drinking whisky; and this latter mode was generally most agreeable to their masters. A slave who would work during the holidays was thought by his master undeserving of holidays. There was in this simple act of continued work an accusation against slaves, and a slave could not help thinking that if he made three dollars during the holidays he might make three hundred during the year. Not to be drunk during the holidays was disgraceful.

1. Why was "this latter mode...most agreeable to their masters" (lines 8)?

(A) It permitted the slaves to return to their work with renewed vigor and interest.
(B) It invigorated the entire plantation with a spirit of well-being and cooperation.
(C) It put to use materials and assets that were difficult to sell on the open market.
(D) It appeared to provide a necessary break in a life of continuous labor and service.
(E) It seemed to confirm the slave owner's belief that slaves were not interested in living as industrious freemen.

2. The tone of the last sentence is

(A) ironic and bitter
(B) jovial and hilarious
(C) pedantic and learned
(D) servile and cooperative
(E) cajoling and pleading

3. The passage is from an autobiographical account by

(A) James Baldwin
(B) LeRoi Jones
(C) Frederick Douglass
(D) Richard Wright
(E) Ralph Ellison

Commentary

This passage raises several questions about its appropriateness in a literature test.

- The passage is about slavery, a highly emotional subject for some test takers.
- The passage is written by Frederick Douglass, an important figure in Black history.
- The passage presents a picture of the kind of subtle influences a master used to control the behavior of slaves.
- The passage specifically mentions drunkenness among slaves.

Considering these issues, some reviewers would deem the passage inappropriate because, although a factually accurate account, it is affectively negative. Others would consider the passage appropriate to use, in accordance with the views expressed by the panel of historians, in a test designed for history students. Resolving the issue of whether the passage should be included in a given test will require considerable discussion among an extended group of people, all involved either in the development of the test or in the sensitivity review process. No matter what their decision about the passage, however, the items contain options that are inappropriate. The test taker who chooses to answer the first item with option B, for example, has clearly misread the passage or has some insupportable ideas about slavery. This test taker, however, has no opportunity to discover that B is an incorrect answer. It is to avoid reinforcing such ideas in the minds of those who choose the wrong answer that options like B are to be removed from test items, in accordance with the ETS sensitivity review guidelines. Rewording other options in the item, particularly A and E, will improve the item from the sensitivity reviewer's point of view. Option E, for example, should have a word like *erroneous* or *mistaken* inserted before *belief*, and *freemen* should be revised. Similar changes in wording are called for in item 2. Options B, D, and E convey an impression of the writer's attitude that is not appropriate.

Example

Questions 3, 4, and 5 refer to the following excerpt from a United States Supreme Court decision:

"That woman's physical structure and the performance of maternal functions place her at a disadvantage in the struggle for subsistence is obvious. This is especially true when the burdens of motherhood are upon her. Even when they are not, by abundant testimony of the medical fraternity, continuance for a long time on her feet at work, repeating this from day to day, tends to injurious effects upon the body, and, as healthy mothers are essential to vigorous offspring, the physical well-being of woman becomes an object of public interest and care in order to preserve the strength and vigor of the race."

3. The views expressed in the excerpt above most probably supported legislation

(A) regulating the hours and working conditions of women
(B) prohibiting the employment of women in specified industries
(C) providing medical clinics for women in specified industries
(D) encouraging the use of birth-control techniques
(E) permitting health insurance companies to charge higher rates to women employed in specified industries

4. In arriving at these views, the Supreme Court

- **(A) followed a strict constructionist line**
- **(B) held as closely as possible to precedent**
- **(C) admitted the legal relevance of statistical, sociological, and historical data**
- **(D) paid close attention to the intent of the legislature**
- **(E) followed the Dillon rule**

5. The views expressed in the excerpt above most closely reflected those of contemporary

- **(A) socialists**
- **(B) feminists**
- **(C) Progressives**
- **(D) eugenicists**
- **(E) Democrats**

Commentary

From the point of view of the sensitivity reviewer, this stimulus is unacceptable for use in an ETS test. It restricts women to their traditional roles as mothers and wives and makes no provision for the student who, unaware of the source of the quotation, accepts the view presented as an accurate one espoused by ETS. Further, the head note, by pointing to the Supreme Court as the source of the quotation, gives stature to the point of view expressed.

From the point of view of the historian, the cited decision is a crucial one in United States history, in that it represented a departure in the method by which the Supreme Court justified the decision (as question 4 indicates). It is also important in the history of women's rights because it defined the legal status of women (a woman is not a person under the law) and led to legalized discrimination against women, but eventually gave impetus to the movement to give women equal rights by constitutional amendment. Further, the decision upheld protective labor legislation for women that provided for such things as stools for saleswomen to sit on so that they would not have to stand for 12 hours a day. Humane treatment for women workers eventually led to humane treatment for all workers, and the decision is therefore crucial in the history of the American labor movement.

When two such equally important issues are at stake—a view of women that might upset some candidates and reinforce stereotypical thinking in others versus a need to preserve the integrity of the subject matter of American history, unpleasant though it may be—the need for compromise is apparent. In this case, a brief statement identifying for the test taker the date of and historical context for the decision will make the quotation acceptable *in a history test*. Such a compromise might have little effect in some other situations on items related to the stimulus, the inclusion of a date definitely changes the nature of the task required by the questions, particularly question 5. That question, from the test assembler's point of view, may have to be omitted from the test. When the constraints of the sensitivity guidelines and the constraints of the pool of items available to meet test content specifications cause such opposing tensions, working out a mutually acceptable compromise is never easy, but it can usually be done.

Mr. **Edwards**. Thank you very much, Dr. Dwyer, and Miss Rigol. There is a vote on the floor of the House of Representatives. We will recess for about ten minutes.

Mrs. **Schroeder**. Mr. Chairman, could I just say one thing? I'm not going to be able to come back because I have to work on this bill. But I just want to say that I'm very disappointed because, when I hear you saying that women are more accurately predicted by the tests than men, then you're really saying we're overpredicting men. That is the whole basis of what the first panel was saying—why are we overpredicting men?

The second thing that disturbs me is that you are telling us that there is this new group of women taking the test—but not to worry, that they're from a lower socio-economic scale I don't think that's relevant. I think it is what kind of academic backgrounds they have. My understanding is the new group of women, and the fact that more are taking the tests, is because they're being encouraged to by their advisors because of their academic performance. So that's what I think is relevant.

If they are really academically much lower than the males taking the tests from high school, then that's different. But I don't care about their socioeconomic background as much as I do their academic background. If you could address those two things for the record, I want the statistics on what kind of academic background they have, not their income background.

That is the second bell and we do have to run. I'm sorry, Mr. Chairman.

Mr. **Edwards**. That's all right. We have plenty of time.

Dr. **Dwyer**. Let me just say—and Miss Rigol may certainly want to comment, too—that the women who take the SAT now are less likely than the men, to have taken the academic program in high school, for example. I think that speaks to the level of their academic preparation. We all have a sense of what the academic program is. There are more women than men, who haven't taken that program, who take the SAT.

[Whereupon, the subcommittee was in recess.]

Mr. **Edwards**. The subcommittee will come to order.

Well, Dr. Dwyer and Miss Rigol, is it your testimony that your testing products, your examinations, are just about as good as they can be made, insofar as bias is concerned, that they are very fair?

Ms. **Rigol**. I believe they are. I don't believe that we should be complacent and say we're never going to look again and they are perfect, just to put them on the shelf. I think we have to continue to evaluate them.

But based on what we know so far, I believe they're as fair as they possibly can be and we'll just continue to work at it. That's my response.

Dr. **Dwyer**. I would agree with that. I would also be immodest and say that I think the test development procedures used at ETS set a standard for other test developing organizations. But I would get a stage more nitty-gritty than that and say that not only do we not have to become complacent, but we make new tests every day. We have to apply the procedures that we do have every day. In doing so, I think we continually learn better ways to do it.

The test sensitivity review standards that I entered into the record, for example, just within the past couple of months were completely revised because we felt that our reviewing experience had taught us so much about how to do that process that we needed to rethink it.

Mr. **Edwards**. So what you're saying is there was some bias in the past, but you think you have eliminated it?

Ms. **Rigol**. I would like to respond.

I think that society has changed. There are things that would have been considered perfectly acceptable by the vast majority of us 20 years ago that grate on us now. In many cases, these are subtle changes. Looking back over old SAT items, I notice things—for example, we might have had a math question about how many shirts can a woman iron in three hours. Well, that is just not reflecting real life situations any more—for many women, at any rate. Those kinds of questions have been changed.

So, obviously, we have to keep updating——

Mr. **Edwards**. So that's obviously bias?

Ms. **Rigol**. If we had that in there now. I'm not sure if it was biased 20 years ago or not. But certainly our perceptions of what——

Mr. **Edwards**. Well, we had a pretty biased world 20 years ago, so far as women were concerned, so you certainly do know it was biased. Of course, it was. Twenty years ago we weren't interested in having women involve themselves to the extent that they are involved in American life, so that was bias.

Ms. **Rigol**. That would be bias.

Mr. **Edwards**. How do you look for bias? How do you find it?

Dr. **Dwyer**. I start off by, first of all, looking at the individual test questions in terms of what it is they're supposed to be asking. I realize this sounds like I'm starting a long way back. But I think the kind of items that invite different interpretations from men and women or from blacks and whites are poorly written or confusing items. So I think a good safeguard against bias is to make sure that you know exactly what educational point you intend to test and checking very carefully to make sure that it doesn't drag in a lot of superfluous information that might be related to factors like race and sex.

The other things that I mentioned earlier get at very specific kinds of bias, the bias that can be introduced by having a person read something in a testing situation that they find upsetting or offensive or that triggers in their mind some response that is just not productive and not related to the educational goals of the testing.

Mr. **Edwards**. Insofar as math is concerned, young women coming from rigorous schools, prep schools—for example, Concord Academy or places like that—do they do just as well as young men coming from rigorous prep schools?

Ms. **Rigol**. I don't know offhand, but I would be glad to gather the information for a group of some selective and rigorous independent schools and compare men's and women's scores. I can provide that to the subcommittee.

Dr. **Dwyer**. May I be permitted a personal anecdote?

Mr. **Edwards**. Yes, please.

Dr. **Dwyer**. My daughter is a college student, a mathematics major, and was shocked to find, when she looked into the number of graduating math majors at Harvard, in the year she was looking at, I believe there were ten people in math, of whom only two were women. That is not the overall sex ratio there.

Mr. **Edwards**. Thank you.

Ms. LeRoy.

Ms. **LeRoy**. Are there not studies that control for these factors, that show that when you discount for socioeconomic status and when you discount for educational background, when you take girls and boys who have had the same number of years of math and come from the same kinds of backgrounds, that girls still perform worse than boys on these tests?

Ms. **Rigol**. No, I think the data is just the opposite, that when you adjust, or at least take into account, the academic preparation of the groups, that the gaps are not nearly as wide.

Ms. **LeRoy**. But they're still there.

Ms. **Rigol**. I again do not have that data right in front of me, but I do know that, generally, when you control for number of years of academic study, you do account for a large proportion of the difference.

Ms. **LeRoy**. ETS has done those studies, is that what you're saying, and that that's your understanding of their results?

Ms. **Rigol**. The College Board has——

Ms. **LeRoy**. I'm sorry, I meant the College Board.

Ms. **Rigol** [continuing]. Has displayed the information on a number of years of study in a number of publications, and will be coming out with a series of reports this next fall that will show, by specific course preparation, the scores for different groups. It does show that it accounts for at least some—I don't now if it is all—of the gap.

Ms. **LeRoy**. We would be interested in seeing those studies when they're done.

Ms. **Rigol**. Yes, I can certainly provide that information, too.

Ms. **LeRoy**. Are there some types of questions, math questions, for example, that you know women do as well on or better than men, and are there some types of questions that you know that men do better on than women? And what is the ratio of those types of questions within the SAT?

I mean, have either of your organizations thought about including more of one type of question to try to adjust the test score differential?

Dr. **Dwyer**. Let me try to address that.

In most testing situations, you don't start with the premise that you want to construct a test that makes groups equal. You start with an educational premise of some sort. I think, as a generalization, across a number of tests, you would be looking at either the distribution of courses that material is to be based on, or within a course of the topics to be covered. The specifications for the test, the determination of how many of what kind of questions to be included, should follow from those educational considerations rather than from score difference considerations.

There have been and the earliest study I can quickly think of is maybe 1958, people doing research with very small samples of items and people, that look at particular variations of questions—for example, setting a question in a male context and then in a female context, and seeing what the sex differences are. There have been some mixed findings on that, and I think part of that is a disagreement about what ought to be considered "male" and what ought to be considered "female." One of the original studies called a question about sewing the "female" version, and a question about a snail crawling up the wall the "male" version. That makes us think it's a little hard to interpret.

We have generally observed that females tend to score lower than males in math and science areas and higher in humanities and the arts, when other things are equal. Again, I'm speaking very broadly over a lot of different kind of data.

Ms. LeRoy. What about within the math area? That's really the focus of my question. For example, data sufficiency questions as opposed to spatial relationship questions, one group does better than the other, right?

Dr. Dwyer. Yes. There is a large body of psychological literature that examines the question of sex differences in spatial relations. That is probably a much better researched area than any of the others that I've spoken of.

Ms. LeRoy. And do the College Board and ETS look at those tests to see how those questions—you know, what percentage of the test is made up of different types of questions?

Dr. Dwyer. Oh, yes. For all the tests that we produce, we begin with a set of test specifications that describe how many of what kind of items go into the test.

Ms. LeRoy. And how do you come out on those kinds of—I mean, what is the ratio? Do you have those statistics?

Dr. Dwyer. Let's see. You know, I can picture a chart in my mind this morning, and I'm not sure if I'm going to be able to read it from my mental image.

I'm not going to be able to do that accurately.

Ms. LeRoy. Well, could you provide that information to the subcommittee, if it's available, at a later time?

Dr. Dwyer. You're thinking about the percentage of——

Ms. LeRoy. The types of questions that you know that there are sex differences on in terms of the correctness of the answer or the number of correct answers. Women answer certain kinds of questions—have an easier time with certain kinds of questions, and men have an easier time with other kinds of questions. How are the tests made up in terms of how much of each of those types of questions they include?

That's a very unscientific description of what I'm getting at.

Dr. Dwyer. I know what you're getting at. I think what I can say is that I can tell you exactly what's on the test, and I can show you also the research that speaks to differences between men and women on things that are like what's on the test. But I can't form a causal relationship for you.

Are you talking about the SATs?

Ms. LeRoy. Yes.

Dr. Dwyer, in an article that you wrote in 1976, you said the ETS method for deciding the content of the SAT systematically favors boys and that—this is a quote—"probably an unconscious form of sexism underlies this pattern. When girls show superior performance, balancing is required; when boys show superior performance, no adjustments are necessary."

Have you changed your mind about that?

Dr. **Dwyer**. I have not changed my mind about the basic statements that I made at that time, that I believe unconscious sexism causes people to accept, without question, women's inferior performance. I believe that my comments, and those of many other people in the same vein, have led to a number of differences in the way that we address bias at ETS and elsewhere.

I would also like to say that, at the time that I wrote that, there was virtually no interest in women's mathematical scores, and that situation has since been dramatically reversed. In the research field, and not just at ETS, but throughout educational research, that is a topic of intense interest these days.

Mr. **Edwards**. Counsel?

Mr. **Slobodin**. Thank you, Mr. Chairman.

Democratic counsel was interested in a study on the question of math performance and what percentage of that is accounted for by the preparatory aspect of it.

From the study I had quoted earlier, which is quoted in Miss Rosser's report—or is cited as a source in her report—the same study, it says "Based on a more broadly representative sample, their study suggests—" this is a Pallas and Alexander 1983 study "—suggests the male/female gap in SAT mathematical performance shrinks considerably when differences in quantitative high school course work is taken into consideration. They report that when differences in quantitative course work were controlled in their analyses, the male/female gap in SAT mathematical performance decreased from 35 points to 14 points."

Then further on down it says, "A recent study using data drawn from the 1977 through 1978 National Assessment of Education Progress in Mathematics found that 25 percent of the variance in mathematics achievement was accounted for by background variables, such as characteristics of the community and educational level of parents, while the number of semesters of mathematics studied explained an additional 34 percent of the variance."

So I would ask the panelists, putting that data together, what gap do we have at this point? First of all, are you familiar with those studies, and what kind of gap do we have at this point, once you take into account the level of preparation and other kinds of background variables that are in that study?

Dr. **Dwyer**. The most significant gap to me is the fact that we have those differences in the level of preparation and participation in mathematics—and I want to throw in science there, too, as a plug. I mean, I think that's where the really problematic issue is, that girls either through socialization are opting out of that stream, or, if they had any original interest in it, they are being funneled away from it.

Since you are interested in that topic, I should probably mention that there is another body of research that speaks to that point. I think the most prominent researcher there is Dr. Elizabeth Fennema, who has studied sex differences in mathematics extensively, beginning around the mid-seventies. Her work looks at, among other things, differential treatment and reinforcement of men and women within their math classes. Her thesis is that it is not just the taking of the math course itself but what happens to you once you get in that course. That is something else that should be considered.

Mr. SLOBODIN. Let's break things down a bit, because when you break down under certain groups, you can get certain disparities, and then you look at it differently. Either the disparity grows or it gets smaller.

Is there a disparity, by sex, in terms of the high test score performance? Have you seen a difference in the decrease in test scores above 600 between male and female, looking at it at that breakdown?

Ms. RIGOL. There have been shifts, and unfortunately I cannot recall that data right now. I would like to send that, also.

Mr. SLOBODIN. How about below 300, looking at it from the same standpoint.

I would also want to follow up on the majority counsel's request, that when you look at the types of questions where there's a real disparity between the correct response rate between women and men, that you also look at race. Because I'm wondering, as you try to neutralize, could it possibly be we're playing with a Rubik's cube here? We may solve the problem of eliminating a question that has a disparity in terms of sex, but it may actually increase the problem in terms of race.

Dr. DWYER. Your analogy of the Rubik's cube is a very apt one, from where I sit.

One of the problems that the balancing question encounters is that the patterns do not run the same way in sex comparisons and in race comparisons. Additionally, they do not necessarily run the same way when you're doing various minority group comparisons. It is sort of hard to know what to do when you have a question that favors Blacks and disfavors Hispanics, for example.

Mr. SLOBODIN. OK. And just on this under- and overprediction issue, is it your opinion that if there were separate sex equations in terms of the predictions of first-year academic performance, that it would not only eliminate the underprediction of females but the overprediction for males?

Dr. DWYER. Yes. Separate sex equations, if you're predicting—This is such a hard topic to talk about, it's so technical. I think, though, it would be fair, given my level of expertise on this subject, for me to say that when you have separate sex equations, you virtually eliminate any question of under- or overprediction on the average. I mean, you have set up a target here and you're sending an arrow straight for it, rather than trying to make a compromise between the two targets that are far apart.

Mr. SLOBODIN. Just one last question. On the underprediction, I am looking here at this table. They break it down by majors. It happens that this table looked at—they conclude, first of all, looking at it as a whole, the difference ratios observed in this study must be considered very small. But in the discipline studies, where there were differences, the highest underprediction was 28 percent of a GPA standard deviation in an engineering program, made up 5 to 1 of men.

Do you have a feel that in courses a lot of these things are skewed as a result of this kind of thing? For example, I would be interested to see what would happen if you took that engineering out or some of the extremes and looked at courses where there was equal participation rates. There may be something here at work.

Have you seen that there's a different interest level, a difference based on sex, an interest in things like computers?

Dr. **Dwyer**. Oh, absolutely. You know, women and men do take different courses and have different majors.

I think it was interesting, that it was mentioned earlier about LSAT and its prediction of law school grades. I mean, that's one of the few instances where you have essentially a situation where men and women take all the same courses in the first year and underprediction is not a problem in that situation as it is at colleges. There you are predicting to everyone's grades together and the men and women have taken systematically different courses.

Mr. **Slobodin**. Thank you.

Ms. **LeRoy**. I would like to ask both of you just one question on test use.

Both of your organizations have guidelines and policies for the proper use of these—well, let's just talk about the SAT. In fact, one of the things you submitted for the record is the ETS Standards for Quality and Fairness. On page 24 it says "ETS will set forth clearly to all score recipients principles of proper use of tests," et cetera, et cetera; "ETS will establish procedures by which fair and appropriate test use can be promoted and misuse can be discouraged or eliminated." Later on it talks about investigating complaints or allegations of improper use, to consult with the sponsor to determine whether to continue services, et cetera.

I guess my question is what actually happens in terms of implementing these guidelines. For example, I think every witness here—whatever they feel about these tests—has said they should never be used as the sole criterion for any decision making purpose. I take it you agree with that?

Dr. **Dwyer**. Yes.

Ms. **Rigol**. Yes.

Ms. **LeRoy**. Well, the Empire State scholarships obviously focused only on SAT scores. There are certain programs run by Johns Hopkins for which I think junior high school children are selected based on their performance on the SAT math score. From what you have said, I would assume that you think those may be improper uses of those tests.

Why do you provide those tests to those people for those purposes if they are improper?

Ms. **Rigol**. There are several answers to that, and I'll try to be brief.

One is that, indeed, the College Board does not support the use of scores alone, unless there is just no other way to do what it is you want to do. We certainly do tell all of our users about the guidelines. When we are aware that a use is being made of the scores that is not appropriate, regional staff—and the College Board does have staff throughout the country to work with institutions—they generally do talk to the organization, or whatever it is, to try to suggest appropriate uses.

There are a number of things that are used in the gifted and talented programs. There is documentation that shows, for their initial selection purposes, when they want to do a national search and try to find as many talented people from all over the country, that the use of the SAT does work very well. There are other criteria that are taken into consideration in the selection of the students who participate in those programs.

I think that part of your question was why do we continue to provide the tests. Well, the tests are provided to students and the students choose to send their scores where they wish to. And while we have debated whether or not we should tell students "yes, you may take the SAT, but we will not send your scores where you would like to have them sent" would not be a proper interpretation of——

Ms. LeRoy. That's sort of a "Catch 22", though, isn't it? You can't get a scholarship in New York unless you take the test, so the student says "I'm not going to take the test because I know this is an improper use of the test", but it's the only way he's going to get the scholarship.

Ms. Rigol. In the case of New York, I think there's an even more important thing, and that is one related to social values and whether the intent of the scholarship program is to recognize scholastic ability or whatever factor, regardless of the composition of the population competing for those awards, or whether the awards should be intentionally subdivided between various subpopulations.

It is not at all unusual—and this is the case in some States and scholarship programs—for the scholarship to be actually set up where a certain number of scholarships must come from certain congressional districts, or you could even say we have to have an equal number of men and women, or you could say we want to have the scholarships come from a certain proportion of various ethnic or monority groups.

No single test is going to do all of that. You have to determine what it is you want to accomplish and then set the guidelines accordingly.

Ms. LeRoy. Have you ever not provided the test to someone based on failure to follow your guidelines?

Ms. Rigol. ETS has, yes. I can't think of a single case right now for the SAT, but ETS has.

Dr. Dwyer. Yes. There is something that I think is important to know about that. We have refused to provide services where we feel that the client using those services is engaged in a misuse of the services, and where frankly the client has just proved intractable in moving towards a better use.

Fortunately—and I say fortunately on purpose—those situations are few in number. There are many questionable testing activities that ETS simply does not engage in from the beginning. But in situations where we are already engaged in a testing activity and a misuse comes to light, we try very, very hard to correct that misuse, rather than to write off the clients. But when that becomes necessary, we do it.

I think our president has been very forthright in his commitment to continue maintaining standards that way. It was he who has instituted the implementation of our standards, which are periodically revised. Our programs are periodically audited against those standards. We invite visiting panels of distinguished educators to come and critique ETS on its adherence to those standards.

Ms. LeRoy. Thank you.

Mr. Edwards. Thank you both very much.

Dr. Dwyer. Thank you, sir.

Mr. Edwards. The last panel today will consist of Mr. Michael Behnke, Director of Admissions, Massachusetts Institute of Technology, and Dr. Denise Carty-Bennia, Professor of Law at Northeastern University, and Executive Chair of Fair Test at Boston.

Mr. Behnke, you are first. We welcome you both. Your full statements, of course, will be made a part of the record. You may proceed. We apologize for keeping you waiting so long.

Do you solemnly swear or affirm that the testimony you are about to give is the truth, the whole truth, and nothing but the truth?

Mr. Behnke. I do.

Dr. Carty-Bennia. Yes.

STATEMENTS OF MICHAEL C. BEHNKE, DIRECTOR OF ADMISSIONS, MASSACHUSETTS INSTITUTE OF TECHNOLOGY; AND DENISE CARTY-BENNIA, PROFESSOR OF LAW, NORTHEASTERN UNIVERSITY, AND EXECUTIVE CHAIR, FAIR TEST, BOSTON, MA

Mr. Behnke. Thank you, Mr. Chairman.

I have been asked to discuss how standardized tests are used in the admissions process at colleges with competitive admissions and to describe research we are doing at the Massachusetts Institute of Technology on gender and testing.

My own experience has been that standardized tests are used in a responsible manner in the process for which they were designed—that is, college admissions. A central question is whether students are denied admission to the college of their choice because of test scores. The fact is that most students are admitted to their first choice college. According to the Higher Education Research Institute, 71 percent of freshmen are enrolled in their first choice college; 93 percent are enrolled in their first or second choice college.

There are relatively few colleges and universities which deny admission to many of their applicants. I have worked in admissions at three colleges—Amherst College, Tufts University, and MIT—which accept fewer than one-third of their applicants. In these cases, it is certainly true that many students are denied the opportunity of enrolling in their first choice college. The admission decision at these institutions, however, is based on many factors and rarely, if ever, is dependent solely on test scores.

Mr. Edwards. If you don't mind my interrupting, while you're on that point, you said that 71 percent of applicants get their first choice, and yet at MIT and places like that, only a third of them do, 33 percent; is that correct?

When you say 71 percent, you're referring to the entire pool of applicants?

Mr. **Behnke**. I'm referring to the entire pool of applicants in the United States. At MIT this year we have admitted about 25 percent of our applicants.

Mr. **Edwards**. That would apply to Stanford, Harvard and Yale?

Mr. **Behnke**. It would be, if anything, a lower percentage.

I have submitted with my written testimony a profile which describes our selection procedure. The process is a complex one and is highly subjective. No one has come close to perfecting the art of human assessment. We can't measure motivation, curiosity, determination, or other similar qualities, but evidence of these qualities can carry great weight. Academic achievement is crucial, but we are faced with trying to understand the meaning of grades and courses from thousands of secondary schools, many of which in recent years have abandoned class rank, which makes it very difficult for us to judge the quality of grades within a particular school.

Extracurricular accomplishments are important, but we must decide whether a leadership position listed on the application means that someone won a popularity contest with little follow through or really served as a positive force for change.

We also find that advantaged students seem to be able to present themselves much more effectively than students without access to good counseling or parents who have gone to college. There is a lot of room for bias in the presentation of extracurricular accomplishments.

This complicated process of judging evidence operates under pressures from many quarters, including alumni, coaches, politicians and donors. In the midst of all this, tests provide a useful standardized measure. They are accepted by professionals as one more very imperfect measure of potential and are used in combination with all other pieces of evidence. The result of this is a very wide distribution of scores.

I have to apologize, as I have an incorrect figure in the written testimony. At MIT last year, the number of applicants who scored above 750 on the SAT math was actually 2,224, not 1,913, as indicated in my written testimony.

Since we admitted only 1,750 applicants, we could have restricted ourselves entirely to this group. Instead, we admitted 40 percent of them. We then admitted 28 percent of those who scored between 700 and 740; 22 percent of those between 650 and 690; 19 percent of those between 600 and 640; 10 percent of those between 550 and 590; 7 percent of those between 500 and 540; and 3 percent of those who scored below 500.

By saying that tests are used responsibly in the admissions process, I don't mean to imply that there is a settled, widely agreed upon way in which tests are used. I mean two things. First, I mean that admissions officers rarely depend on test scores alone to deny someone admission, and second, I mean that admission professionals periodically reexamine their use of test scores, as they do the other criteria they use. This reexamination has, in fact, led some schools to no longer require test scores or to modify their requirements.

Admissions professionals are troubled by test abuse. Testing is a topic of debate at almost every meeting of admissions professionals. All of us, I think, are trying to find out more about them so that we can use them responsibly. I sent two of my staff members to attend the recent meeting of Fair Test. We are troubled by scholarship agencies which use score cutoffs; we are troubled by the use of score cutoffs for athletic eligibility. We are troubled by the use of college entrance exams in identifying gifted students in middle or even elementary schools. When most of the students labelled "gifted" are white males, what message does this send to young women and minorities? We worry about people defining their worth and potential in terms of test scores. This is especially troubling because of race and gender differences in scores. We worry about the growing industry of test coaching schools feeding off people's anxieties. We ultimately worry that we may lose a useful piece of evidence in deciding college admissions because test abuse may lead to abandoning or weakening these tests. We hope that public scrutiny and pressure and the efforts of testing agencies themselves will lead to fewer abuses and more public understanding.

In the meantime, there are some things that we can do in admissions. The first is to communicate to the public how tests are used. The profile I have submitted is an attempt to show students and counselors the range of test scores so they do not focus on averages. I should mention, though, that even the use of that profile is problematical. It's been used at MIT for a number of years and it was first issued in response to concerns about providing information to consumers. But my student/faculty committee on admissions feels that it overemphasizes the use of test scores, as they understand how they're being used in the process of which they are a part. And while people want the information, my faculty committee feels we may even be providing too much information and misleading people in the way that I think Mrs. Schroeder pointed out.

The second thing we can do is research. We have been doing research at MIT to examine the relationship between college grades and an academic prediction formula combining high school grades in math and science, rank in class, if available, and SATs and achievement test scores. The formula has a scale which extends from 1 to 99. We looked at the relationship between that index and grades for a recent entering class, most of whom had graduated. MIT has a graduation rate within five years of approximately 85 percent. The mean academic index on this scale of 1 to 99 for men who entered was 68, and for women it was 60. We looked at what made up the difference and the lower average for women reflected lower test scores. Specifically, there were statistically significant differences in the math SAT, the math achievement test and the science achievement test.

The correlation of that index with grade point average in the senior year was 0.47 for both men and women. I think the point has been made earlier here today that, in fact, the test does predict grades as well for women as it does for men. We have found that to be true. This means that the index is a reasonable predictor of 8-term cumulative average for both men and women; that is, low scoring men and low scoring women tend to receive lower grades at MIT. But the index underpredicts the final grade point average for women at all levels.

Because of the lower index, we would expect women to have a lower GPA. In fact, at the end of eight terms, there was no significant difference in grade point average. It is also useful to note that women have a higher retention rate at MIT, so this is not due to women dropping out at a higher rate.

Various theories have been advanced to explain the fact that standardized tests underpredict the academic performance of women. One is that the result may be due to differences in the selection of courses by women and men. We looked at each individual major at MIT and found the same thing going on within departments, so we found no evidence of course selection affecting this at MIT.

We believe the possibility needs to be explored that the tests do not really get at what female ability means. In the words of our Associate Director of Research, Dr. Elizabeth Johnson, "What we are really suggesting is the notion of multiple models of developed ability, perhaps even intelligence, which are culturally related and which, if accurately drawn, would predict equivalent success on real world outcome measures."

In conclusion, I think it is important to note that we were able to do this research at MIT because for many years MIT has admitted women with somewhat lower test scores. This happened because the admissions committee did, in fact, look at many factors, including grades and extracurricular activities and whatever kinds of personal qualities we could identify, and based on the entire picture, we felt that the women being admitted were every bit as strong as the men, in spite of somewhat lower test scores, and in fact that has turned out to be true. So the research has not surprised us.

Thank you, Mr. Chairman.

[The statement of Michael C. Behnke, with attachment, follows:]

Testimony Presented to the Subcommittee on
Civil and Constitutional Rights, April 23, 1987

Michael C. Behnke
Director of Admissions
MIT

I have been asked to discuss how standardized tests are used in the admissions process at colleges with competitive admissions and to describe research we are doing at MIT on gender and testing.

My experience has been that standardized tests are used in a responsible manner in the process for which they were designed, i.e. college admissions. A central question is whether students are denied admission to the college of their choice because of test scores. The fact is that most students are admitted to their first choice college. According to the Higher Education Research Institute, 71% of freshmen are enrolled in their 1st choice college; 93% are enrolled in their 1st or 2nd choice.

There are relatively few colleges which deny admission to many of their applicants. I have worked in admissions at three colleges - Amherst, Tufts and MIT - which accept fewer than one-third of their applicants. In these cases, many students are denied the opportunity of enrolling in their first choice college. The admission decision is based on many factors and rarely is dependent solely on test scores. I have submitted with my written testimony a profile which describes our selection procedure. The process is a complex one and is highly subjective. No one has come close to perfecting the art of human assessment. We can't measure motivation, curiosity, determination or other similar qualities, but evidence of these qualities can carry great weight. Academic achievement is crucial, but we are faced with trying to understand the meaning of grades and courses from thousands of secondary schools. Extracurricular accomplishments are important, but we must decide whether a leadership position listed on the application means that someone won a popularity contest with little follow through or really served as a positive force for change.

This complicated process of judging evidence operates under pressures from many quarters including alumni, coaches, politicians and donors. In the midst of all this, tests provide a useful standardized measure. They are accepted by professionals as one more very imperfect measure of potential and are used in combination with all the other pieces of evidence. The result of this is a wide distribution of scores. At MIT last year, we had 2,224 ~~1915~~ applicants who scored over 750 on the SAT - Math. Since we admitted only 1750 applicants, we could have restricted ourselves to this group. Instead we admitted 40% of them. We then admitted 28% of those who scored between 700 and 740, 22% of those between 650 and 690, 19% of those between 600 and 640, 10% of those between 550 and 590, 7% of those between 500 and 540 and 3% of those below 500.

By saying that tests are used responsibly in the admissions process, I don't mean to imply that there is a settled, widely agreed upon way in which tests are used. I mean two things. First, I mean that admissions officers rarely depend on test scores alone to deny someone admission. Second, I mean that admission professionals periodically reexamine their use of test scores, as they do other criteria. This reexamination has led some schools to no longer require test scores or to modify their requirements.

Admission professionals are troubled by test abuse. Testing is a topic of concerned debate at almost every meeting of admission professionals. [I sent two of my staff members to attend the recent meeting of Fair Test.] We are troubled by scholarship agencies which use score cut-offs. We are troubled by the use of score cut-offs for athletic eligibility. We are troubled by the use of college entrance exams in identifying "gifted" students in middle or even elementary school. When most of the students labelled "gifted" are white males, what message does this send to young women and minorities? We worry about people defining their worth and potential in terms of test scores. This is especially troubling because of race and gender differences in scores. We worry about the growing industry of test coaching schools feeding off people's anxieties. We worry that we may lose a useful piece of evidence in deciding college admissions because test abuse may lead to abandoning or weakening tests. We hope that public scrutiny and pressure and the efforts of testing agencies themselves will lead to fewer abuses and more public understanding.

In the meantime, there are two things which college admission professionals can do. The first is to communicate to the public how tests are used. Our profile is an attempt to show students and counsellors the range of test scores so they do not focus on averages. The second is to do research.

We have been doing research to examine the relationship between college grades and an academic prediction formula combining high school grades in math and science, rank in class if available, and SAT's and Achievement Test scores. The formula has a scale which extends from 1 to 99. We looked at the relationship between the index and grades for a recent entering class, most of whom had graduated. (MIT has a graduation rate within 5 years of approximately 85%). The mean academic index for men who entered was 68; for the women it was 60. The lower average for women reflects lower test scores. Specifically, it reflects statistically significant differences in Math SAT, the Math Achievement Test and the Science Achievement Test.

The correlation of the index with grade point average (GPA) in the senior year was .47 for both the men and the women. This means that the index is a reasonable predictor of 8-term cumulative average for both men and women, i.e. low scoring men and low scoring women tend to receive lower grades. But the index underpredicts the final GPA for the women at all levels.

Because of the lower index, we would expect women to have a lower GPA. In fact, at the end of 8 terms, there was no significant difference in GPA. It is also useful to note that women had a higher retention rate.

Various theories have been advanced to explain the fact that standardized tests underpredict the academic performance of women. One is that the result may be due to differences in the selection of courses by women and men. We found no evidence of that at MIT. We believe the possibility needs to be explored that the tests do not really get at what female ability means. In the words of our Associate Director for Research, Dr. Elizabeth Johnson, "What we are really suggesting is the notion of multiple models of developed ability, perhaps even intelligence, which are culturally related and which if accurately drawn, would predict equivalent success on real world outcome measures."

Summary of Testimony Presented to the Subcommittee on Civil and Constitutional Rights, April 23, 1987

Michael C. Behnke
Director of Admissions
MIT

I have been asked to discuss how standardized tests are used in the admissions process at colleges with competitive admissions and to describe research we are doing at MIT on gender and testing.

My experience has been that standardized tests are used in a responsible manner in the process for which they were designed, i.e. college admissions. An admission decision is based on many factors and rarely is dependent solely on test scores. The process is a complex one and is highly subjective. The result of this is a wide distribution of test scores in any entering class.

Admission professionals are troubled by test abuse. Admission tests are being used for many purposes other than college admission. We hope that public scrutiny and pressure and the efforts of testing agencies themselves will lead to fewer abuses and more public understanding.

In the meantime, there are two things which college admission professionals can do. The first is to communicate to the public how tests are used. The second is to do research.

We have been doing research to examine the relationship between college grades and an academic prediction formula combining high school grades in math and science, rank in class if available, and SAT's and Achievement Test scores. We found that women had significantly lower scores on the Math SAT, and the Math and Science Achievement tests. The index was a reasonable predictor of 8-term cumulative grade point average (GPA) for both men and women. But the index underpredicted the GPA for women at all levels. Whereas we would have expected a lower GPA for women, there was no significant difference between women and men.

Michael Behnke came to MIT as Director of Admissions in May of 1985. For nine years previous to that he was Dean of Admissions at Tufts University. At Amherst College between 1971 and 1976, he served as Associate Dean of Admissions, Dean of Freshmen and a lecturer in American Studies. Mr. Behnke has taught in both public and private secondary schools, in the Upward Bound Program for low income students, and in the Peace Corps in Sierra Leone, West Africa. He was also the Education Director for a community action agency in Springfield, Massachusetts.

Mr. Behnke received his undergraduate degree in American Studies from Amherst College and a Master of Arts degree in the same field from Penn. He presently serves as Chairman of the New England Regional Council of the College Board, Vice-Chairman of the National Advisory Committee on International Education, and as a member of the scholarship selection committees of the National Merit Corporation and United Technologies Corporation.

Massachusetts Institute of Technology

Office of Admissions • Room 3-108 • Cambridge, MA 02139
• Phone (617) 253-4791 • Telex 92-1473

MESSAGE FROM THE DIRECTOR

A Different and Exciting Class

We made some special efforts this year to tell prospective applicants about the broad choices available at MIT both in the curriculum and in student life. In response, applications increased by 8 percent including a 15 percent increase in women. This larger pool allowed us to admit a more diverse group of students with the same academic talent necessary to succeed at MIT. This year's class is 38 percent women and 30 percent minority students (189 Asian Americans, 61 Black Americans, 23 Mexican Americans, 17 Puerto Ricans and 4 Native Americans). We also had a decrease in the number of students interested in Electrical Engineering (our largest major) and an increase in interest in such fields as economics, management, political science and humanities.

A New Selection Procedure

During the year we examined the selection procedure used at MIT for many years. While the procedure has served MIT well, we decided it was time for a change. We will continue to have applications evaluated by two people, with a third reader called in to resolve any significant differences. Readers will continue to be drawn from the staff, faculty and administration. But the ratings they give will be changed. Instead of one academic rating based primarily on grades and test scores in math and science, applicants will receive two academic ratings. One will be similar to the old rating in that it will be a numerical summary of grades and test scores. More attention, however, will be given to a student's whole record rather than primarily the math and science record, and more attention will be given to the quality of courses taken. A second academic rating will be completely subjective. It will be a reader's impression of an applicant's personal characteristics pertinent to academic promise. We hope to recognize students who bring a special level of excitement to the classroom or an unusual brilliance to their own studies or research.

There will also be two personal ratings. One will measure a student's actual accomplishments and skills. This may be talent in music or athletics, expertise in a hobby, leadership, or entrepreneurship. It can also simply recognize that a student has been limited in this regard by the necessity to work long hours for pay. The second rating will be a subjective reaction to the applicant's individual style and sense of purpose.

We hope that this somewhat more complex procedure will allow the Admissions Committee to make decisions based on more dimensions of the applicants. We think the effect might be to place somewhat more weight on grades and quality of program as opposed to standardized testing and somewhat less weight on small differences in objective measures in favor of trying to recognize real love of learning and other special personal qualities.

Special Initiative for Underrepresented Minority Students

Although the number of black, Mexican American, Puerto Rican and Native American students entering MIT is high compared to most schools, it has not increased in many years. We are concerned by the apparent drop in the number of minority students going on to college and by the increasing anxiety over high cost and over loans.

Wishing to make known our commitment to increasing the number of underrepresented minority students at MIT, we have developed a new service and combined it with several existing ones in something we call the Pathway to the Future Program. The centerpiece is our new service, The Practical Experience Program. We recognize that one of the greatest benefits MIT students enjoy is access to summer jobs which provide a high enough salary to significantly reduce loan burdens. A new person in our Office of Career Services will seek out summer jobs for our underrepresented minority students and counsel the students in how to qualify for those jobs. The opportunities will be in a wide variety of fields such as business, planning, engineering and banking.

photo Garfinkel, *The Tech*

For minority students especially interested in engineering, we will continue our Second Summer Program. This gives MIT underrepresented minority students an opportunity to spend three summer months, at the end of their first year, working in the design or research department of a major corporation. A substantial number of the participants are able to negotiate jobs at the same company in succeeding summers.

MIT invites all underrepresented minority students to Project Interphase which takes place during the summer before freshman year. Students choose between a seven week session which includes freshman courses and a two week session with seminars about freshman courses and a focus on resources outside the classroom. MIT provides room and board as well as all necessary texts and materials for classes.

MIT also provides special financial support to underrepresented minority students who need financial aid. There is some evidence that many minority families have to pay more for basic necessities. Consequently, MIT reduces the usual Parental Contribution for lower-income minority families. Further, if minority students find their academic program necessarily lengthened beyond the usual four years to a bachelor's degree, MIT will provide financial aid, up to need, for a ninth or tenth term of study.

Best Wishes,

Michael C. Behnke
Director of Admissions

BEST COPY AVAILABLE

GENERAL STATISTICS ON THE CLASS

Applications for Admission	US	International
Preliminary Applications	13,131	1,218
Final Applications	5,513	700
Offers of Admission	1,659	103
Expected to Register	928	69

Applicants for Early Action

The Early Action Program exists only for citizens and permanent residents of the U.S.

Early action consideration is available to applicants who have completed the MIT application process by November 1—test scores from the November testing date will be accepted. Early Action applications will be reviewed by the end of December. Some offers of admission will be made; other applications will be held, without prejudice, for consideration at the regular time. Applicants admitted in December need not reply until the Candidates' Reply Date in early May.

	Total
Number Who Applied	1,056
Number Admitted Early	428
Number Deferred and Admitted Later	76

Schools Represented in the Entering Class	Number of Schools	Number of Students
US Public Schools	603	742
US Independent or Church Related	159	185
International	63	70

RANK IN CLASS

Although in each case the secondary school record is a significant predictor of college performance, school standards and marking systems vary so widely that average grades in school cannot be satisfactorily generalized here. the rank-in-class data are less affected by marking systems, but they do not recognize differences in school standards. Therefore, both rank in class and the percentage of the class that goes on to college are considered. Each school is urged to explain how class rank is determined and how grades are affected by accelerated, enriched, or non-grade programs.

Class rank for all applicants for whom rank was submitted

Rank in High School Class	Number of Applicants	Percent Admitted	Number in Class
Top tenth of class	4,123	34%	803
2nd tenth of class	638	7%	25
3rd tenth of class	197	3%	5
4th tenth of class	95	2%	2
5th tenth of class	33	3%	1
Lower half of class	42	0%	0

GEOGRAPHIC DISTRIBUTION

This class represents 34 foreign countries and 48 states. No geographical quotas are imposed on the selection of students from within the U.S.

Region	Number of Applicants	Percent Admitted	Number Enrolled	Percent of Entering Class
New England	907	28%	161	16%
Middle Atlantic	1,540	31%	278	28%
South Atlantic	703	29%	117	12%
North Central	948	30%	151	15%
South Central	441	28%	70	7%
West	994	29%	137	14%
Guam, Puerto Rico & Virgin Islands	57	40%	14	1%
Canada	62	26%	6	1%
International (excluding Canada)	561	14%	63	6%
Total	6,213	28%	997	100%

CREDIT AT ENTRANCE FOR CLASS ENTERING SEPTEMBER 1985*

Studies beyond the level of the traditional secondary school curriculum are recognized for credit and, where appropriate, placement to entering first-year students.

Category	Students Seeking Credit	Students Receiving Credit
Credit by College Board AP Tests	657	562
Credit by MIT Advanced Standing Exam	118	46
Credit by College Transcript	111	101
Credit from 'A' Level Exams and the International Baccalaureate	18	17

*Some students are included in more than one category.

TRAN[illegible]

MIT has the [illegible]

[illegible]

Last year, [illegible]

[illegible]

COLLEGE BOARD TEST SCORES*

This summary of College Board scores indicates: 1) the number of applicants in each interval who completed all application procedures for admission to MIT in 1986, 2) the percent who were actually offered admission, and 3) the number expected to register in the class of 1986. The figures on this page include all applications, U.S., Canadian, and international, on which action was taken, although they may omit a few applicants not coming directly from secondary schools. Our experience over many years reaffirms that standardized tests are important: 1) in comparing the achievement of students from various school systems, and 2) in predicting which of our applicants are most likely to experience academic success at MIT. As can be seen from the figures, no "cutoff" scores of any sort are used. For example, although less than half of the students with mathematics aptitude scores of over 700 were offered admission, a total of 368 students with lower mathematics SAT scores were admitted. There are, of course, some practical limits below which there would be serious doubts about a student's ability to be successful in the freshman year at MIT. By examining the table below, it is possible to get a rough estimate of the probability of admission to MIT for a student with given scores. Overall, we regard the tests as powerful instruments in our search for talent. However, each case is decided on its individual merits: a score of 800 does not ensure admission nor does a score below 600 ensure rejection.

SCHOLASTIC APTITUDE TEST

ACHIEVEMENT TESTS

	Verbal			**Math**			**Math Level I**			**Math Level II**			
Range of Scores	# of Applicants	% Admitted	# in Class	# of Applicants	% Admitted	# in Class	# of Applicants	% Admitted	# in Class	# of Applicants	% Admitted	# in Class	Range of Scores
750-800	154	61	29	1,913	40	400	465	43	111	2,193	39	463	750-800
700-740	669	46	142	2,187	28	364	771	34	157	1,000	29	188	700-740
650-690	1,227	37	235	1,081	22	155	703	23	90	608	24	89	650-690
600-640	1,388	32	264	533	19	55	516	21	73	271	19	29	600-640
550-590	1,058	23	158	201	10	13	247	13	18	115	15	11	550-590
500-540	708	17	92	97	7	7	133	12	12	20	10	1	500-540
Below 500	874	10	75	66	3	1	86	1	0	13	0	0	Below 500

ACHIEVEMENT TESTS

	English Composition or History			**Chemistry**			**Biology**			**Physics**			
Range of Scores	# of Applicants	% Admitted	# in Class	# of Applicants	% Admitted	# in Class	# of Applicants	% Admitted	# in Class	# of Applicants	% Admitted	# in Class	Range of Scores
750-800	195	63	42	512	48	112	93	61	20	596	35	107	750-800
700-740	746	50	180	617	39	136	277	49	59	522	32	99	700-740
660-690	1,201	39	257	611	32	115	317	45	83	520	28	98	650-690
600-640	1,173	30	206	582	31	107	233	28	39	439	19	56	600-640
550-590	985	22	141	393	23	56	158	25	18	312	21	49	550-590
500-540	926	17	96	253	16	26	84	16	11	231	20	31	500-540
Below 500	1,185	9	75	224	11	14	90	9	7	153	9	7	Below 500

---

* Three Achievement Tests are required for application to MIT. The U.S. applicant must choose from each of three groups:
(1) Level I or Level II Mathematics
(2) Chemistry or Physics or Biology
(3) English Composition (with or without essay) or American History and Social Studies or European History and World Cultures

[redacted]

PARTICIPATION IN ACTIVITIES

Members of this class gained admission, at least in part, on the strength of participation in activities such as the following

Activity	Number in Class
Elected school or class officer	187
Varsity sport participant	386
Participant in school publications	469
Officer in school club or organization	397
Officer in civic, community, or religious group	340
Participant in school or community music group	311
Participant in drama, debate, or dance	204
Part-time work during school	542
Full- or Part-time work during the summer	513

In addition to the above, students participated in many other unique activities

photo Campbell

photo Galvin

FIELD
MIT
HOCKEY
7

photo: Poler

POLICY OF NONDISCRIMINATION

MIT admits students of any race, color, sex, religion, or national or ethnic origin to all rights, privileges, programs and activities generally accorded or made available to students at the Institute. It does not discriminate against individuals on the basis of race, color, sex, sexual orientation, religion, handicap, age, or national or ethnic origin in administration of its educational policies, admissions policies, scholarship and loan programs, and other school administered programs, but it may favor U.S. citizens or residents in admissions and financial aid. The Institute has created and implemented and will continue to implement an affirmative action plan expressing its commitment to the principle of equal opportunity in education.

[illegible]

Mr. EDWARDS. Thank you, Mr. Behnke.

We are now going to hear from Dr. Denise Carty-Bennia, who is Professor of Law at Northeastern University, and Executive Chair of Fair Test in Boston.

STATEMENT OF DENISE CARTY-BENNIA

Dr. CARTY-BENNIA. Thank you.

Meaningful access to higher education for racial and ethnic minorities in the country has become stymied almost as soon as it has begun at predominantly white colleges and universities.

This aspect of the American dream realistically can no longer be viewed as simply deferred. In an increasing number of instances, it is being denied. Today, blacks have a smaller presence on American campuses than they did six years ago, both in absolute numbers and as a percentage of all undergraduates. The enrollment of Hispanic students also lags far behind their overall representation in the population.

This situation is most notable in both public and private four-year colleges and universities, as well as graduate schools. In fact, most minorities in higher education are concentrated disproportionately in two-year community colleges, with little or no real possibility that they will transfer to a 4-year institution.

Standardized tests exacerbate this problem. While a number of complex educational, political and social factors contribute to the limited access of minorities to four-year colleges and universities as well as graduate schools, standardized tests continue to be a major factor because of their central and, in fact, growing role in admissions decisions. Many colleges and universities throughout the country are now placing greater reliance on standardized test scores in an effort to so-called "upgrade" their academic standards.

In fact, another recent study of undergraduate admissions by the American Association of Collegiate Registrars and Admissions Officers, and the College Board, found that 39 percent of the public and 42 percent of the private 4-year postsecondary institutions set minimum SAT scores for admission, and that approximately one-third of all 4-year institutions set minimum ACT scores for admission.

These practices should be of great concern to all of us. Automatically rejecting an applicant because he or she did not obtain a minimum score is among the most blatant misuses of standardized admission exams. I will return to that later on.

Test scores systematically hamper the opportunity of minorities to gain admission to American colleges and universities. Racial and ethnic minorities perform poorly on standardized college and admissions tests. Overemphasis on these tests significantly reduces the opportunity of minorities to gain access. I have here a table that I will share with you in my actual report, but it indicates that this is, in fact, supported. Even when family income is considered along with race, it is clear that racial and ethnic minorities do not perform as well as whites at the same economic level on these standardized exams.

Standardized tests are often biased against racial and ethnic minorities. The 1979 New York truth-in-testing law forced test publishers to make public all copies of the university admissions tests.

Their tests recently examined 15 scholastic aptitude tests. Many of the questions in these tests required students to be familiar with the activities and the vocabulary of upper middle-class suburban America. Things such as golfing, tennis, pirouettes, property taxes, kettledrums, minuets, melodeons, timpanists, polo and horseback riding were all mentioned on these tests. Students not familiar with these culturally specific activities could not have obtained the high SAT scores needed to enter America's selective colleges and universities. Nor could they have received financial aid awards from both private as well as governmental agencies.

One example of an SAT question—and I have included several in my report for you—is "melodeon is to organist what reveille is to bugler, solo to accompanist, crescendo to pianist, anthem to choirmaster, and kettledrum to timpanist." I would not have know the answer to this question. It is, in fact, that melodeon is to organist what kettledrum is to timpanist.

Testing for the Public, a nonprofit organization which trains minority students to take standardized tests, recently examined the law school admissions test, the LSAT, using information also made public through New Yorks truth-in-testing law. Their research identified many items which contained degrogatory references to prominent minority figures such as W.E.B. Du Bois, Cesar Chavez, and Harriet Tubman.

They also found numerous items which are extremely offensive to minority test takers. For example, the following recently administered LSAT items were supposed to measure a student's knowledge of grammar:

"Afrikaans is the language of the ruling party in South Africa, of the Afrikaners, whose votes maintain the status quo." No error.

"The Supreme Court ruled that it is not inherently unconstitutional for a white suburb to refuse to change zoning rules which practical effect was to block construction of racially-integrated housing."

Students usually have less than a minute to answer each item on most college and professional school multiple-choice exams. David White, Executive Director of Testing for the Public, points out that such questions often cause minority students to get angry or to waste time thinking about the contents of such questions.

A 1980 study by Joseph Gannon for the National Conference of Black Lawyers provides further evidence of bias in the LSAT. The large gap between the median LSAT scores of blacks and whites historically has been explained away by test publishers as the result of unequal educational opportunity. Grannon's study took care to eliminate the possibility of lower academic ability on the part of minority students as an explanation for his findings. He examined the difference in the LSAT scores of black and white college seniors, from the same universities, and who had earned comparable undergraduate grade point averages. Gannon's finding showed that blacks, with the same grades, from the same colleges as whites, scored more than 100 points lower on the LSAT. In fact, this has been my personal experience with over 10 years worth of law school admissions work and, in particular, work on minority subcommittee admissions process at Northeastern.

Coaching has been talked about earlier today. I maintain that this places minority students in what we call a case of double jeopardy. As the use of the test has increased, a parallel phenomenon has developed—coaching for the test. The SAT alone has spawned a thriving multi-million dollar industry in preparing students to take the test, not to mention the LSAT preparation industry.

Preparaton ranges from private tutors, who charge prices up to $60 per hour, to the coaching schools, like the Stanley H. Kaplan Educational Centers and the Princeton Review, which charge tuition in the $400 to $600 per course range. Both Kaplan and the Princeton Review claim average SAT score improvements in the 150 to 200 point range. Such an increase can mean the difference between a rejection notice and college admission with a scholarship.

This coaching boom, however, puts most minority and low income students in double jeopardy. Not only are they unable to afford the advantages promised by coaching, but the success of coaching increases the disparity in performance between racial groups even further.

Standardized tests are poor indicators of prospective minority students performance. The problems caused by racially disparate scores and standardized test bias are magnified by overreliance on standardized test score admissions. Tests have been oversold and overused to the detriment of other, often better predictors of performance. Many studies, including one conducted in 1985 by Dr. Peter Garcia, Dean of Education at Pan American University in Texas, for the National Institute of Education, have concluded that standardized tests have no predictive ability for future performance. Charles Willy, a professor at the Harvard Graduate School of Education, has reported that at Ivy League colleges there is no correlation between admissions test scores and the academic performance of minority students by their fourth year. Many of these minority students have had to make significant adjustments to the college environment.

Professor Willy also found that when graduate school admissions committees ignore applicant scores on standardized tests, these committees tend to admit a higher proportion of minority students than they do when test scores are made a part of the admissions decision. At the undergraduage level, Willy continues, evidence also exists that the use of scores on standardized aptitude tests as part of the admissions process disproportionately excludes some racial and ethnic minorities.

In 1977, when the University of California proposed a change in its admissions policy that would give greater weight to standardized test scores than to high school grade point averages, a member of the Board of Regents requested that the new criterion be applied hypothetically to the class that had been admitted a year before the new policy was proposed in order to assess the potential effects of the change. The study revealed that if the proposed admissions criteria had been in effect a year earlier, the total University of California student body would have included 9.5 percent fewer Hispanic students and 8.8 percent fewer black students.

Tests are very inaccurate at even what they purport to measure. In the College Board publication "The Admissions Testing Program

Guide for High Schools and Colleges", the Board provides even more detail on the limitations of the scores. The guide urges that SAT scores be interpreted as ranges rather than as points. The ATP guide refers to the fact that the score an individual receives on one administration of the test is probably not the person's true score for an exact measure of that person's ability.

They then go on to speak about what we call normal variations in the score than an individual receives when they are tested on the same or similar test at different times. This means that for most—that is to say, two-thirds—of the individuals taking the SAT, both the verbal and mathematical score obtained will be within 30 points of their so-called true score. If, for example, a student's true verbal score is 450, then the actual score of the student will be between 420 and 480.

The same holds true when we're making distinctions between two persons with respect to their scores. This is called the standard error of deviation or difference. The Board, thus, advises colleges that between students' score differences of less than 66 points and 72 points on the SAT verbal and SAT mathematical, respectively, have little significance. Yet schools often make important judgments on the basis of as little as 10 point differences in the SAT performance. By the way, the same holds true, and is actually more true, for the LSAT. Now that it's rated on a 10 to 48 scale, the kind of distinctions that we make between students are between scores of 36 and 34, two point differences with respect to using the LSAT.

This overemphasis on test scores obviously underemphasizes other and what we would maintain are probably more valid factors, such as grades, extracurricular activities, writing and creative ability.

Standardized tests influence the awarding of scholarship money, which further hampers the ability of minorities to pay for a college education. Over $100 million in merit scholarships are directly influenced by biased standardized tests. Scholarship agencies' reliance on test scores is often the key criteria that keeps minority students from obtaining needed financial aid awards. The National Merit Scholarship Program, for example, automatically rejects students who do not achieve scores in the top one-half or 1 percent in that State from consideration for the $23 million in scholarships awarded annually.

Last year, all recipients of college scholarships awarded by the State of Alabama were white, even though the public school system is nearly 50 percent minority. Information made public through a class action lawsuit revealed that the State's heavy reliance on the ACT scores was responsible for no minority students receiving scholarship assistance. In addition, many colleges and universities scholarship programs award on the basis of standardized test scores.

Minority female high school students are doubly penalized by both the gender and racial biases of the SAT. In every category, males outscore females. This is across all ethnic and racial categories. The New York Empire and Regents Scholarships, worth over $40 million last year, are awarded exclusively on the basis of the student's ACT or SAT score. In 1986-87, this reliance resulted in 72 percent of the winners of the New York Empire Scholarships, worth up to $10,000 each, going to males, and just 28 percent going to females. Bear in mind that's obviously compounded then by the racial discrimination, which we must overlay on these statistics.

Counselors also often rely on test scores as the basis for recommending college and university, not to mention graduate school. My own experience is that minority students who present themselves at pre-law advisors' offices at college and university campuses are regularly discouraged from applying to graduate school programs, and in particular, law school, on the basis of their low performance on the standardized LSAT exam.

In addition, it is fairly clear that students themselves are self-selecting about the schools that they apply to. I wanted to respond to my colleagues on the panel's remarks earlier, around the statistics of those who get into their number one choice of college or university. That has to be viewed against the backdrop that most students readjust the colleges and universities that they apply to, much less the graduate schools that they apply to, on the basis of receipt of their standardized test score. If it is lower, they then place themselves almost automatically opting out of applying to certain "more selective" schools.

Recommendations. It seems to me, at the very least, we ought to be asking the test companies to enforce their own guidelines. If the College Board guidelines on the uses of College Board test scores and related data emphasize that the test score should not be the sole factor in determining the admission of an applicant, then the College Board ought to make sure that the recipients of their College Board scores are, in fact, not only informed, but, in fact, living up to that practice. The College Board continues to provide test scores to colleges and scholarship agencies that automatically deny consideration to students who fail to achieve a certain minimum cut-off score on the PSAT or the SAT. By the way, that's also true for most law schools in this country. They use a minimum LSAT cut-off score. There are two large drawers in a file cabinet in the admissions office at Northeastern University which will not be looked at by the admissions committee because they do not meet a minimum LSAT score.

Similarly, in a statement regarding tests and standards, ETS president Gregory Anrig has stated that admissions test results are supplementary to the academic record and other information about applicants. Test scores should be used in combination with other information and not as the sole basis for important decisions affecting the lives of individuals.

Mr. **Edwards**. If I may interrupt at that point, what do you think would happen at Northeastern if the admissions office accepted a bunch of those applications where the scores were very low?

Dr. **Carty-Bennia**. Probably not a whole lot.

Mr. **Edwards**. What would happen to those students?

Dr. **Carty-Bennia**. I think they would probably graduate from law school and go on to successfully pass the bar and practice law. I'm not suggesting that exceptions are not made. In point of fact, we all know that exceptions are made for students with low standardized test scores. But the basis of those exceptions tends to be on the basis of family connections, alumni connections to the school, and other forms of admitting students who have low standardized test scores. They do manage to graduate and they do manage to pass the bar and they do manage to go on and practice law effectively.

The problem, of course, is that their interest groups are represented in the admissions process, and we're suggesting that that works in a discriminatory way against those who do not have their interest group representative in the office to make that exception for them, with respect to standardized test scores.

The second thing, it seems to me, is that we ought to require that standardized admission tests be made as fair as possible. The Golden Rule bias reduction technique is a safeguard which test companies should employ to ensure that their tests measure relevant knowledge differences between test takers and not irrelevant, culturally specific information. It is based on a November, 1984 out-of-court settlement agreement between the Educational Testing Service, the State of Illinois, and the Golden Rule Insurance Co. of Lawrenceville, IL. ETS agreed to employ this new procedure in order to settle a lawsuit charging that their Illinois insurance licensing exam unfairly discriminated against blacks.

In 1986, ETS extended the Golden Rule reforms to its uniform insurance exam that is annually administered to over 250,000 job applicants in 22 States and Bermuda. The Golden Rule reform makes exams fairer, not easier. Under the procedure, the same content areas are covered as on previous tests, and the exams are of the same level of overall difficulty. The only difference is that within groups of equally difficult items, in the same content areas, test publishers must select those items that display the least difference in the correct answer rates between majority and minority test takers.

While not part of the Golden Rule settlement, the procedure lends itself to the elimination of gender bias in standardized test taking as well. Application of the Golden Rule bias reduction for race and gender bias should be required for every higher education standardized admissions test.

Finally, it seems to me that we ought to talk about opening up the SAT for competitive bid. For the past 40 years, the College Board has simply renewed the contract awarding ETS the right to develop the SAT and related products. Internal ETS documents reveal that ETS earns a profit of about 30 percent from its College Board-related activities. If ETS had competition from this lucrative $50 million a year contract, the company would have to either become more responsive to the concerns noted, or face the very real possibility that a more innovative and responsive company would be awarded the contract.

Finally, I think that we ought to bring to this committee's attention the fact that I just was informed there are certain Federal Government scholarships awarded on the basis of standardized test scores, the Byrd scholarships, which I would ask that this committee at least take the very first steps of looking into and perhaps looking towards eliminating or at least minimizing the impact of standardized tests in the award of those scholarships.

Thank you.

Mr. **Edwards**. Those are good recommendations. Thank you very much.

We're only going to have about five minutes. Why don't you go ahead, Ms. LeRoy.

Ms. **LeRoy**. Mr. Behnke, the subcommittee has heard a lot of testimony today about different types of tests and test uses, but most of the testimony has focused on the SAT as a college admissions test. That already is a small universe, I suspect, in the realm of testing and is, I suppose, a reflection of our own cultural bias as to people in this room as much as anything else. But MIT represents an even more rarified atmosphere, I suppose, even in the world of college admissions.

Do you think that the sort of process that your institution has initiated with respect to college admissions is transferrable to other institutions that are larger, less selective, less elite, I suppose, and can be used with the same kind of reliability as the process that you have instituted at MIT?

Mr. **Behnke**. Well, the quick answer is yes. It's a matter of how many resources the institution is willing to commit to the process of admissions. I think people in the profession want to do as thorough job as possible and they are limited in many cases by staff and other kinds of support. I think as much as possible, decisions on how to allocate resources like a college education should be based on as much information about the individual as possible—the meaning of the grade, the meaning of the secondary schools they come from, the quality of courses they're in, which is something we look at very carefully. Then all the kinds of activities that a person takes part in. That takes some sensitivity and some time in reading applications and getting to know communities, because as I mentioned earlier, I think one of the real problems is the sophistication with which different kinds of people present themselves.

We would very much like to depend more heavily on things like what a student has actually done, what their characteristics might actually be. But a student from an affluent suburban high school or prep school gets recommendations and advice on how to present him or herself substantially different from someone in a small farm community. We're always trying to read into applicants' situations what their context is.

So to go beyond the objective evidence is a time-consuming process and demands resources. Ideally, I think every institution ought to be doing that.

Ms. **LeRoy**. But it is possible to do if they're committed to devoting the time and the resources?

Mr. **Behnke**. Yes.

Ms. **LeRoy**. Thank you.

Mr. **Edwards**. Dr. Carty-Bennia, what do you say to the response of a previous witness, that the make-up of the female population of applicants has much to do with their doing poorer in these tests than they did previously, that there more of them, that they come from poorer families, with less education and so forth?

Dr. **Carty-Bennia**. Well, at least with respect to the LSAT, that isn't true. We still cream, if you will, the "creme de la creme" of students coming from undergraduate institutions who sit to take the LSAT. In fact, what the LSAT does is to underpredict how successful they will be academically in law school. But that's also consistent with the fact that these have been "superstars", if you will, in terms of their academic record in college as well as high school.

Mr. **Edwards**. One quick question of Mr. Behnke. What happens to those applicants that you accept at MIT who really have very low scores?

Mr. **Behnke**. In most cases students with low scores also have low grades, so that looking at all the evidence together, they are not admitted. If the evidence for some reason doesn't match, we in most cases go beyond the evidence. I have gone so far as to call secondary schools and asked them to get teachers out of class to come and talk to me about the student's performance.

That's why the process is very time-consuming, if you're going to go beyond the evidence, and if we're convinced that the test scores don't test that individual's potential, then we ignore them.

Mr. **Edwards**. Counsel for the minority.

Mr. **Slobodin**. First of all, let me just say for the record that I did not get a copy of Professor Carty-Bennia's testimony.

Dr. **Carty-Bennia**. It will be submitted——

Mr. **Slobodin**. It makes it a little difficult to take a look at some of the statements that you made, although a lot of them have been lifted, I think, from the report that Miss Rosser released last week.

You said that instead of standardized tests we ought to look at alternatives like grades and extracurricular activities. I wanted to ask you, what is inherently superior in looking at those items as opposed to the standardized tests? Specifically, I would like you to explain why you think a multiple-choice test in a math class in high school is inherently less bias-free than the math section on the SAT, and how you can say there is less bias in evaluating whether a running back in college, an award-winning running back or someone who played in a symphony orchestra, how can you evaluate—How can you compare apples and oranges and how is that less bias-free than the standardized tests?

Dr. **Carty-Bennia**. First of all, you asked me two questions, so let me try to address the latter part first.

With respect to the way in which law schools can evaluate a running back, if you will, as opposed to someone who plays in an orchestra, it seems to me that schools make very up-front policy decisions about the kind of diversity that they want to see in terms of the composition of classes at their law school. We regularly get applications from a wide variety of people who do not translate, if you will, into being comparable oranges or comparable applies. They are, in fact, apples and oranges. But we are very desirous of having not only apples and oranges but bananas and cherries as well.

Mr. **Slobodin**. You're starting off with a result. You want a certain percentage right at the start. Suppose there aren't enough bananas coming in or——

Dr. **Carty-Bennia**. It's an imperfect world that we live in. In any given class, we probably won't have the appropriate percentage of the target percentage that we were thinking about in any given year. But we strive, out of the people that we get, with any certain comparable pool, to get the best of that pool. And so we will look at all of the running backs or comparable athletes. We will look at all of the comparable orchestra players, and we will look at all of the persons that are from rural areas to some extent and try to get the best out of each of those groups for a mix.

Mr. **Slobodin**. Well, you're explaining the process but you're not telling me why there is less bias in that process——

Mr. **Edwards**. I'm sorry to interrupt, but we have to——

Dr. **Carty-Bennia**. Because they'll be compared on proven past record.

Mr. **Edwards**. I hate being unfair to our witnesses. You really were entitled to a lot more time. But we did run out of time. We've got a supplemental on the floor and there's a vote right now. If I just get over there, I'm sure that education money will be saved. So thank you very much.

Dr. **Carty-Bennia**. Thank you.

[Whereupon, at 12:45 p.m., the subcommittee was adjourned.]

○